Self-Supervised Learning

안인규 (Inkyu An)

Speech And Audio Recognition
(오디오 음성인식)

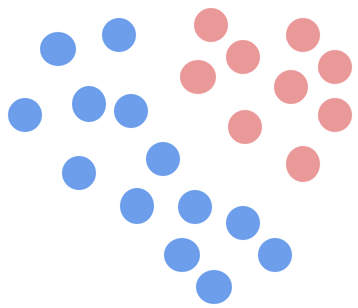
https://mairlab-km.github.io/

This lecture material refers to
https://github.com/yandexdataschool/speech_course?tab=readme-ov-file and
https://github.com/markovka17/dla

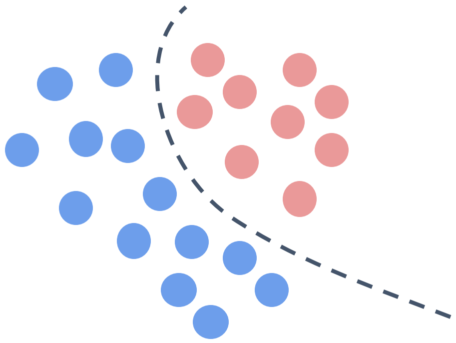
Today lecture …

- Learning Paradigms
- Self-Supervised Learning
- Self-Supervised Learning for Speech
- beyond ASR

Learning Paradigms

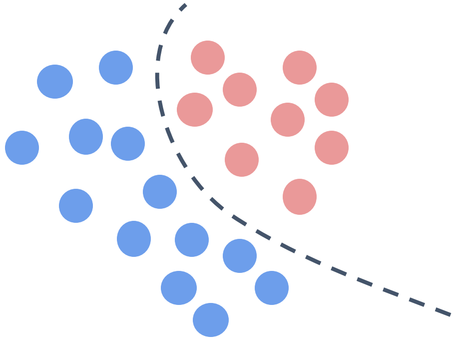


Learning Paradigms

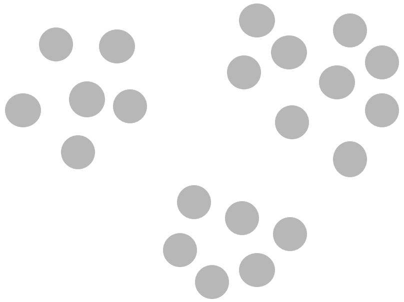


supervised learning

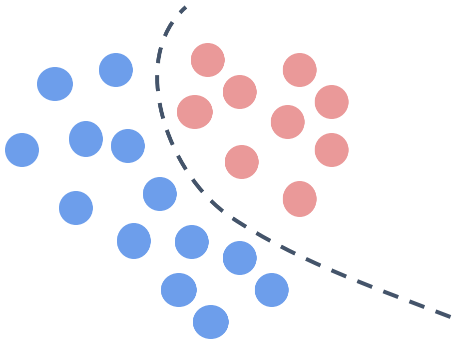
Learning Paradigms



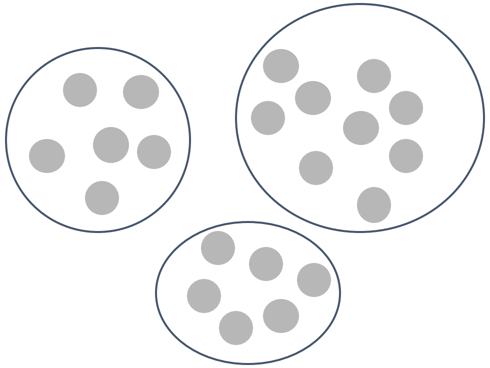
supervised learning



Learning Paradigms

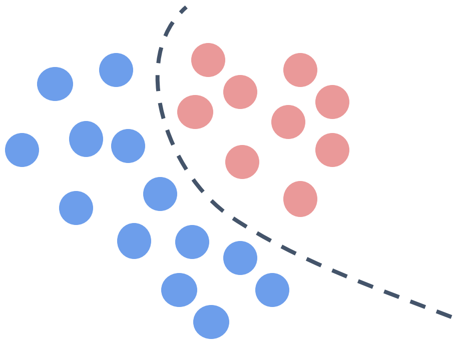


supervised learning

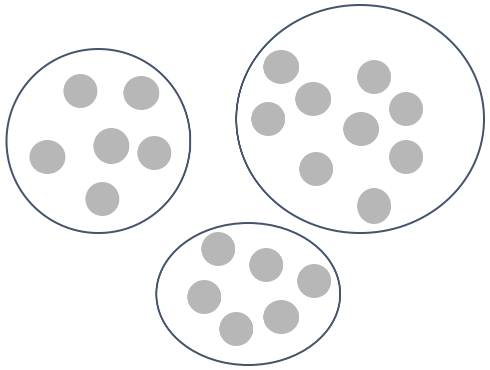


unsupervised learning

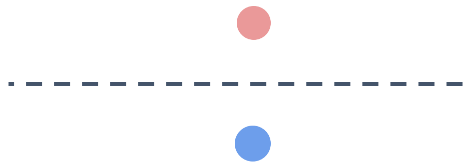
Learning Paradigms



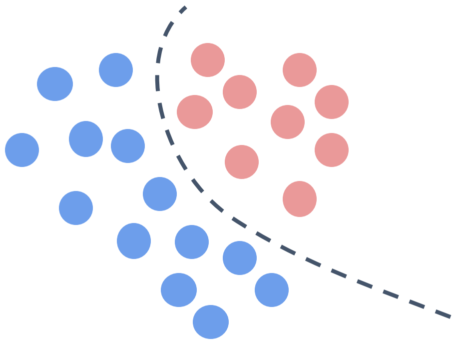
supervised learning



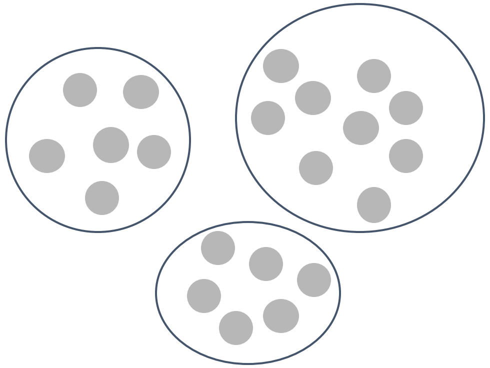
unsupervised learning



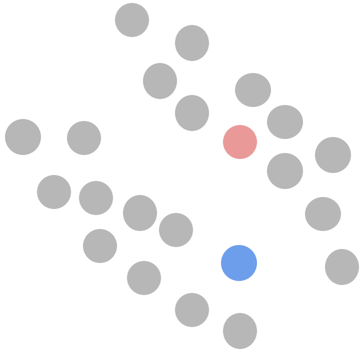
Learning Paradigms



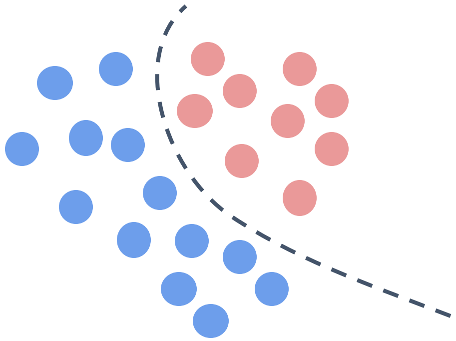
supervised learning



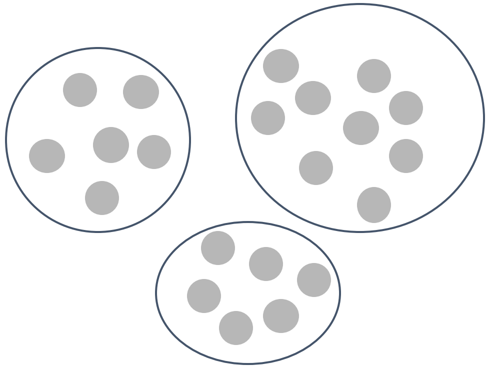
unsupervised learning



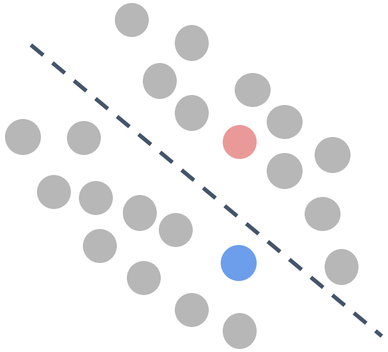
Learning Paradigms



supervised learning

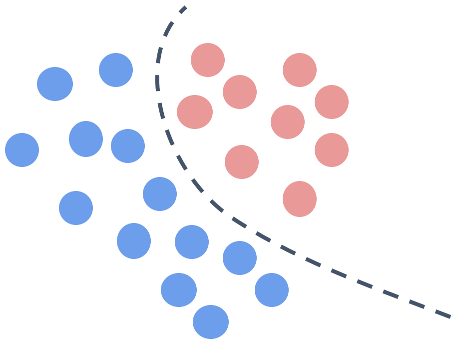


unsupervised learning

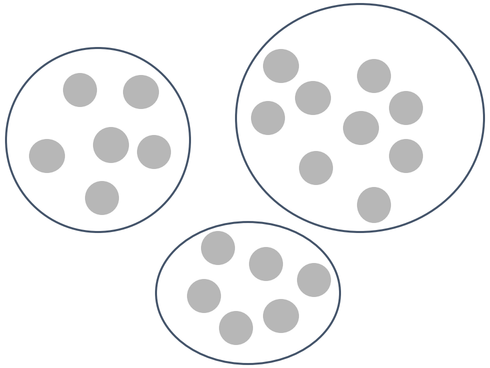


semi-supervised learning

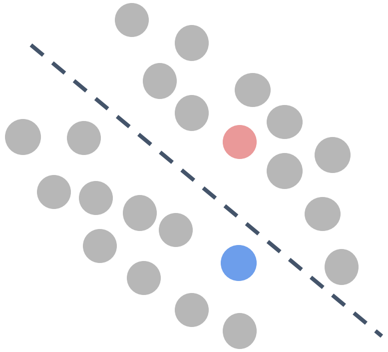
Learning Paradigms



supervised learning



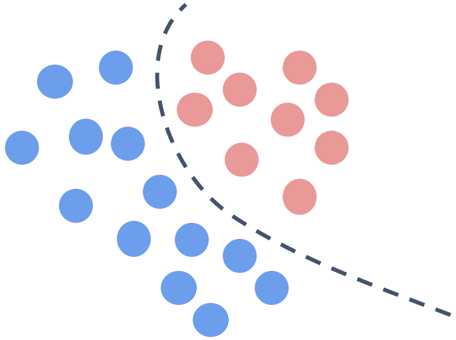
unsupervised learning

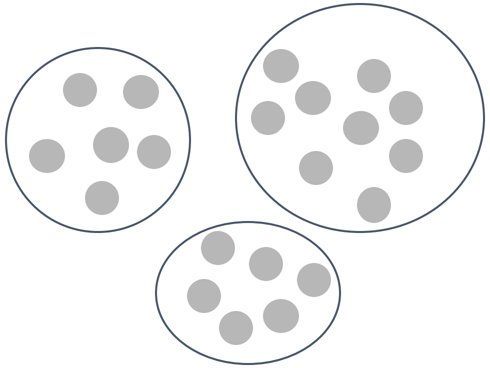


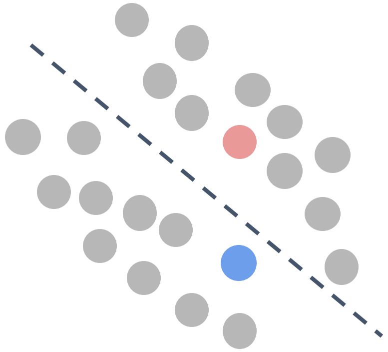
semi-supervised learning

#capybara #pets #animals #wild

Learning Paradigms

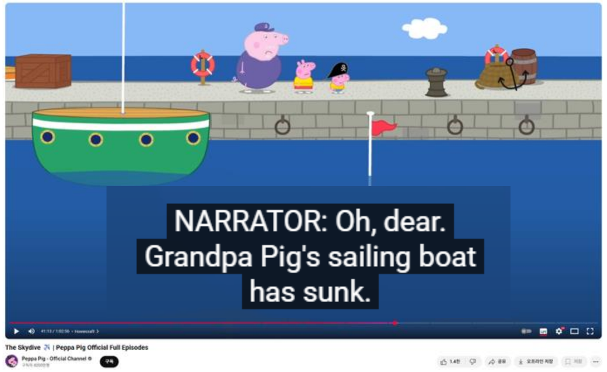

supervised learning


unsupervised learning


semi-supervised learning


#capybara #pets #animals #wild

⇒ hashtag prediction



⇒ speech recognition

weakly-supervised learning

weakly-supervised learning for Speech

- Meta MMS (**M**assively **M**ultilingual **S**peech), 2023
  - Scaling Speech Technology to 1,000+ Languages
  - New Testament
    - faithcomesbyhearing.com
    - goto.bible
    - bible.com
  - Towards Robust Speech Representation Learning for Thousands of Languages (CMU; 2024)

- OpenAI Whisper, 2022
  - Robust Speech Recognition via Large-Scale Weak Supervision
  - 680k hours
  - We construct the dataset from audio that is paired with transcripts on the Internet

Self-Supervised Learning

Image:

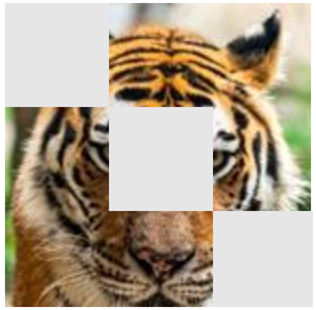
Text:

They studied life
from books, while
he was busy living it.

Self-Supervised Learning

input features

Image:

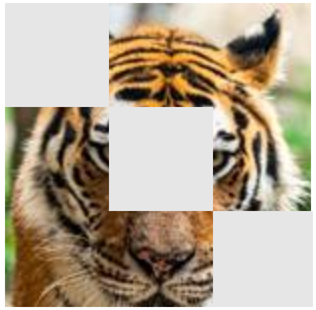


Text:

They studied life from books, while he was busy living it.

They [MASK] life from books, while he was [MASK] living it.

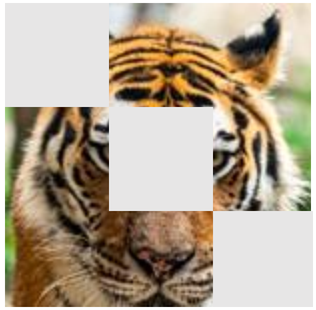
Self-Supervised Learning

		input features	target variables
Image:			
Text:	They studied life from books, while he was busy living it.	They [MASK] life from books, while he was [MASK] living it.	They studied life from books, while he was busy living it.

Self-Supervised Learning

input features target variables

Image:



Text:

They studied life from books, while he was busy living it.

They [MASK] life from books, while he was [MASK] living it.

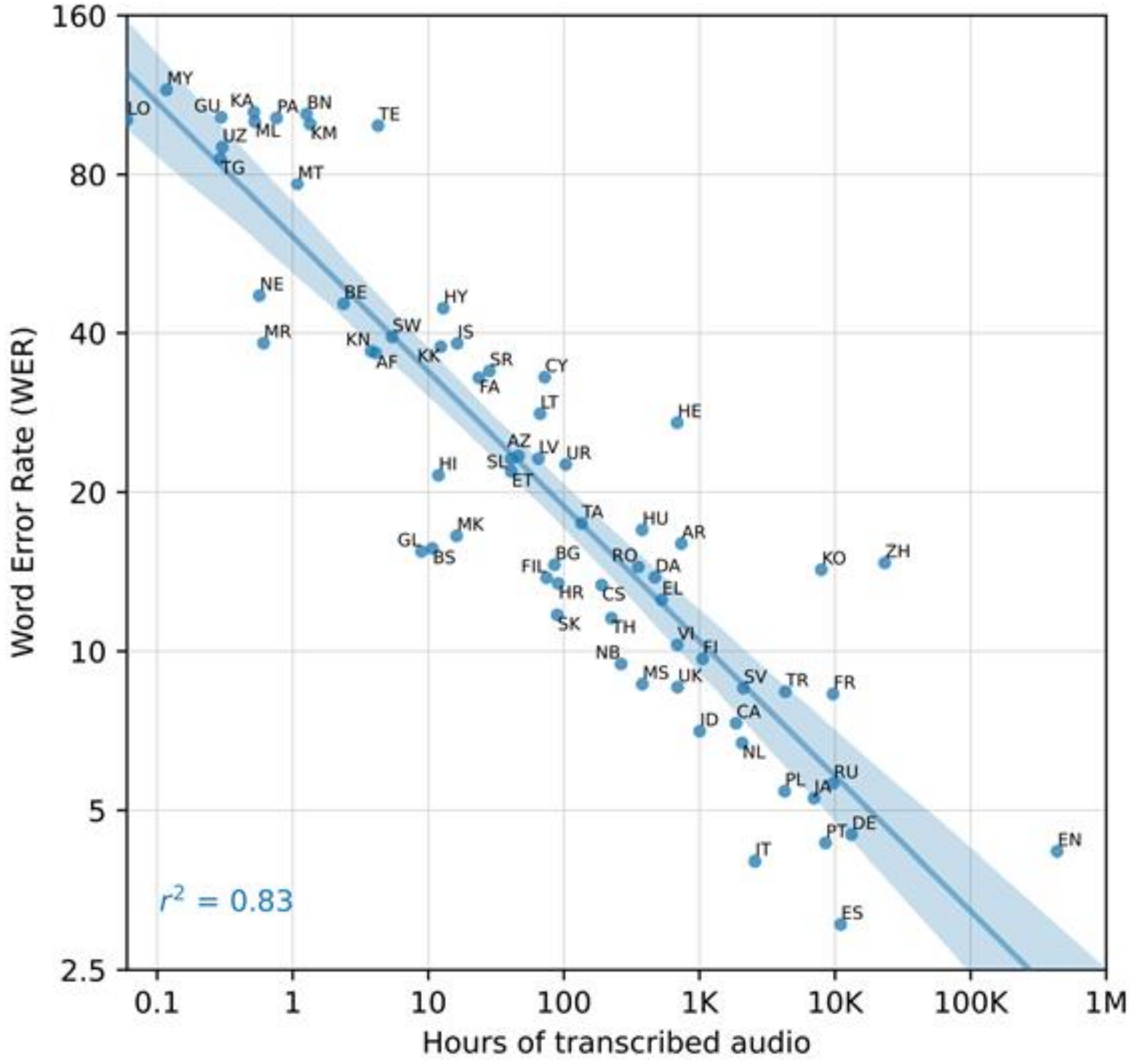
They studied life from books, while he was busy living it.

"In self-supervised learning, the system learns to predict part of its input from other parts of it input. In other words a portion of the input is used as a sup ervisory signal to a predictor fed with the remaining portion of the input."

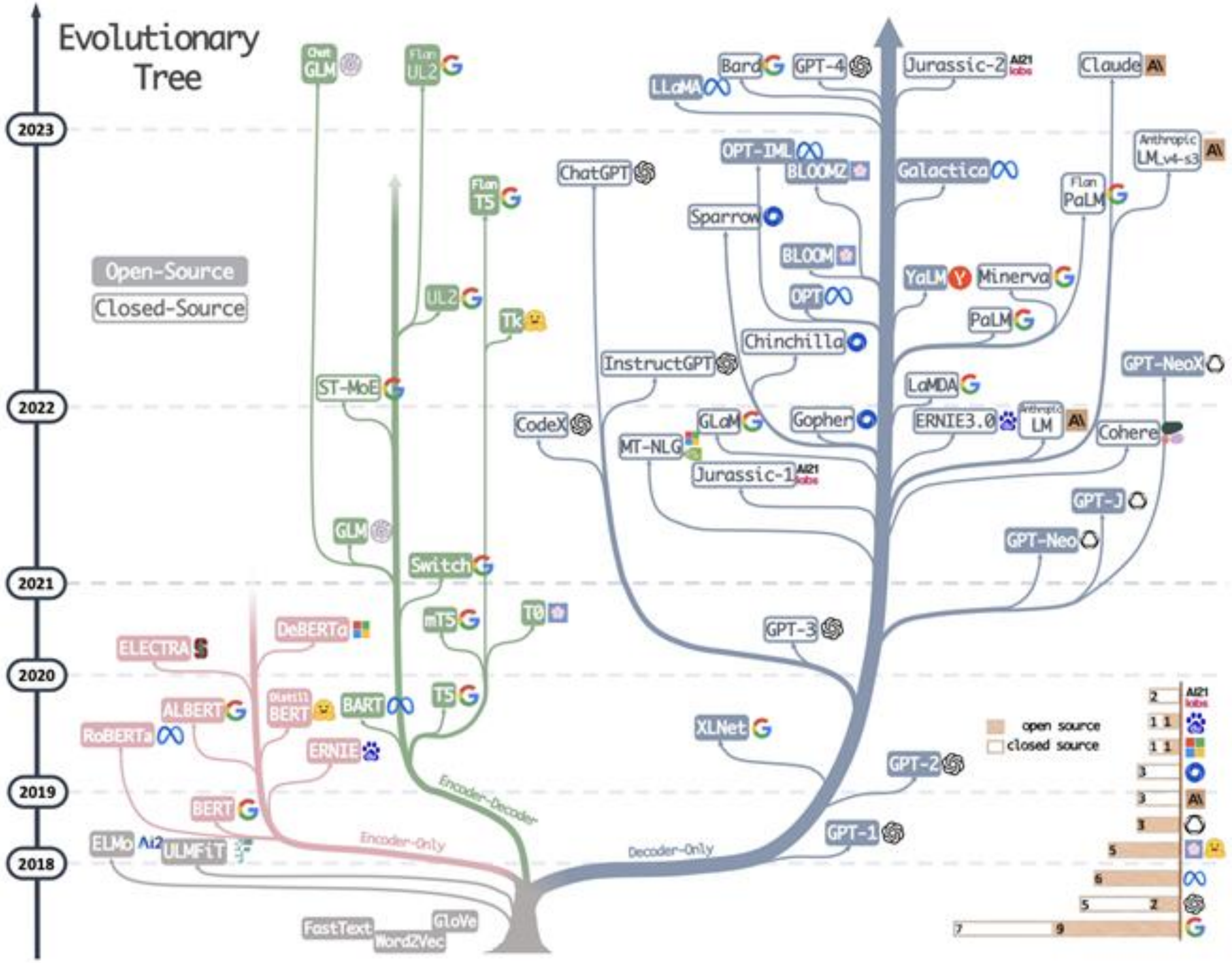
https://www.facebook.com/722677142/posts/10155934004262143/

Self-Supervised Learning

- fast convergence
- better convergence (data)
- low-resource
- large-scale models (data)
- foundation model



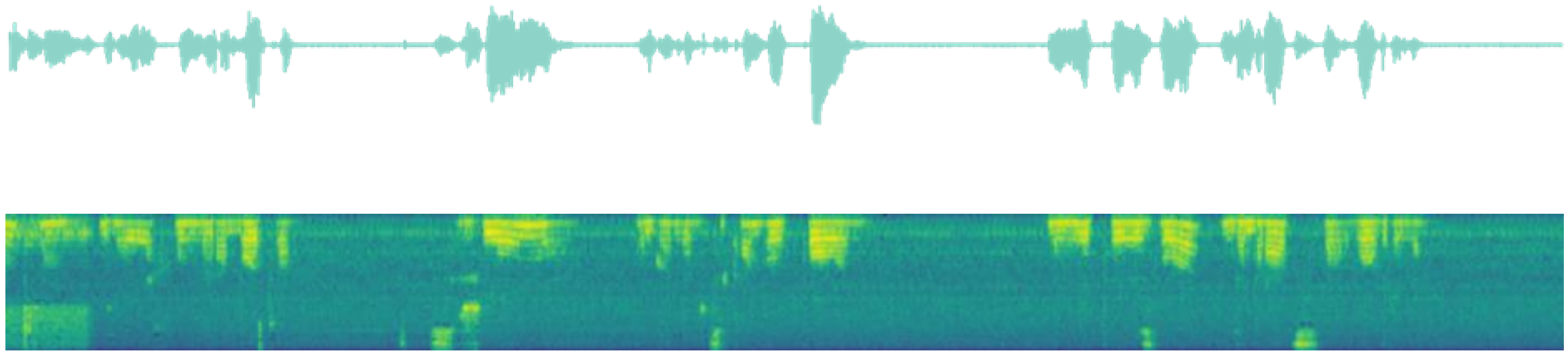
Self-Supervised Learning for Text



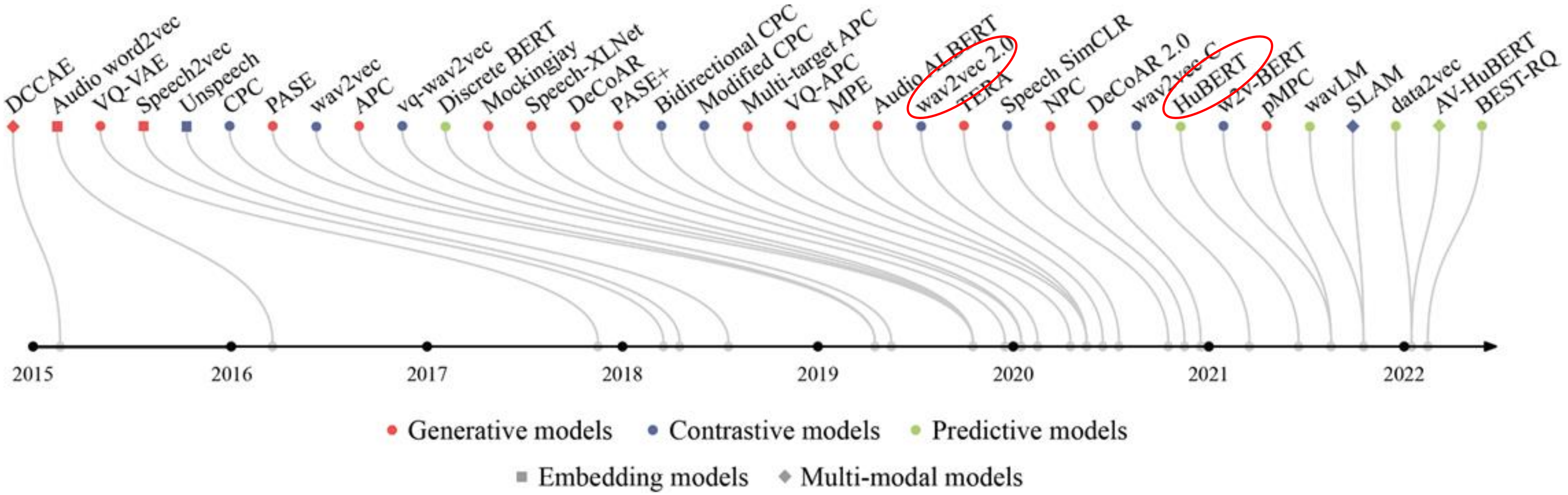
[Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond]()

Self-Supervised Learning for Speech

- 1 speech second ~ 16000 floats
- how to build vocab?

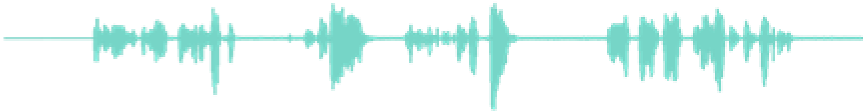


Self-Supervised Learning for Speech

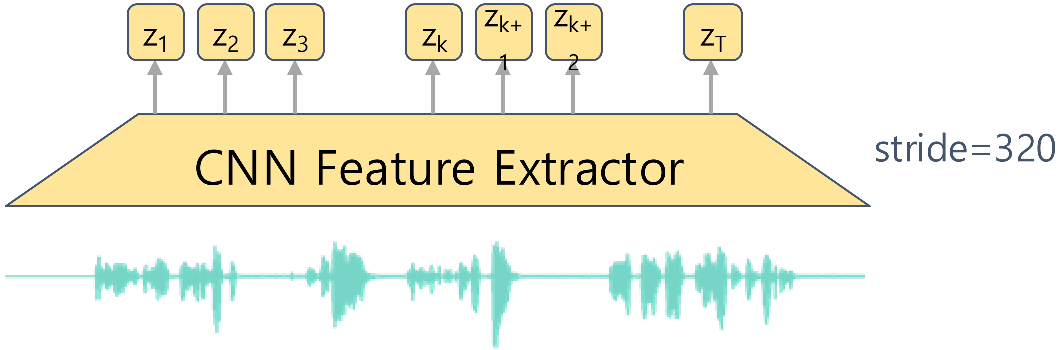


ICASSP 2022 SSL tutorial

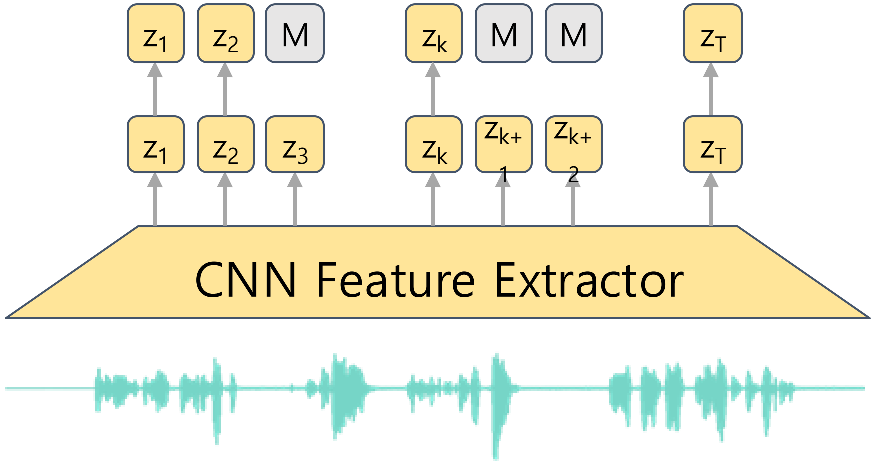
wav2vec2.0



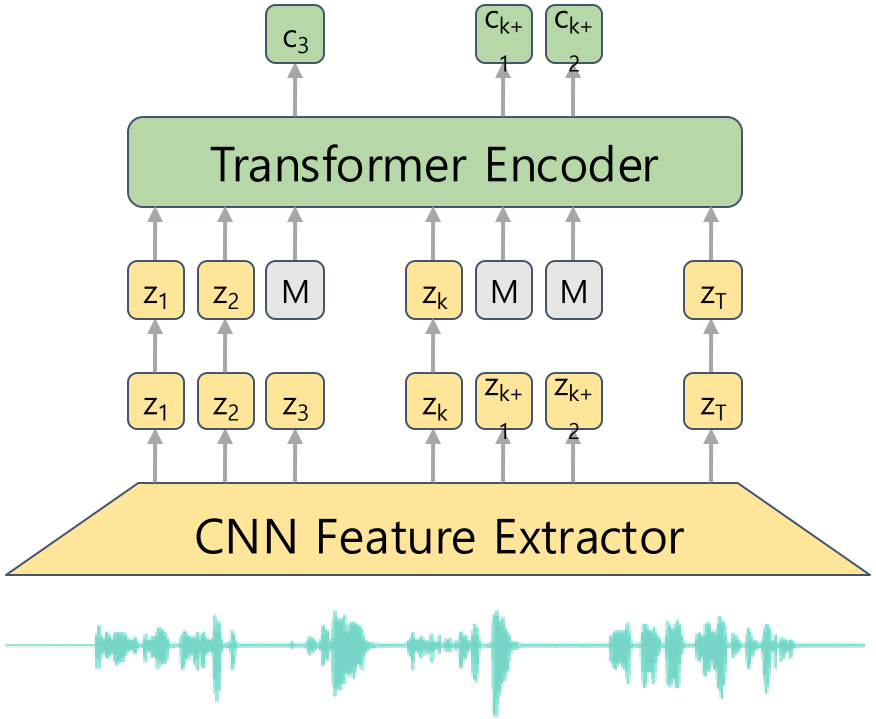
wav2vec2.0



wav2vec2.0



wav2vec2.0

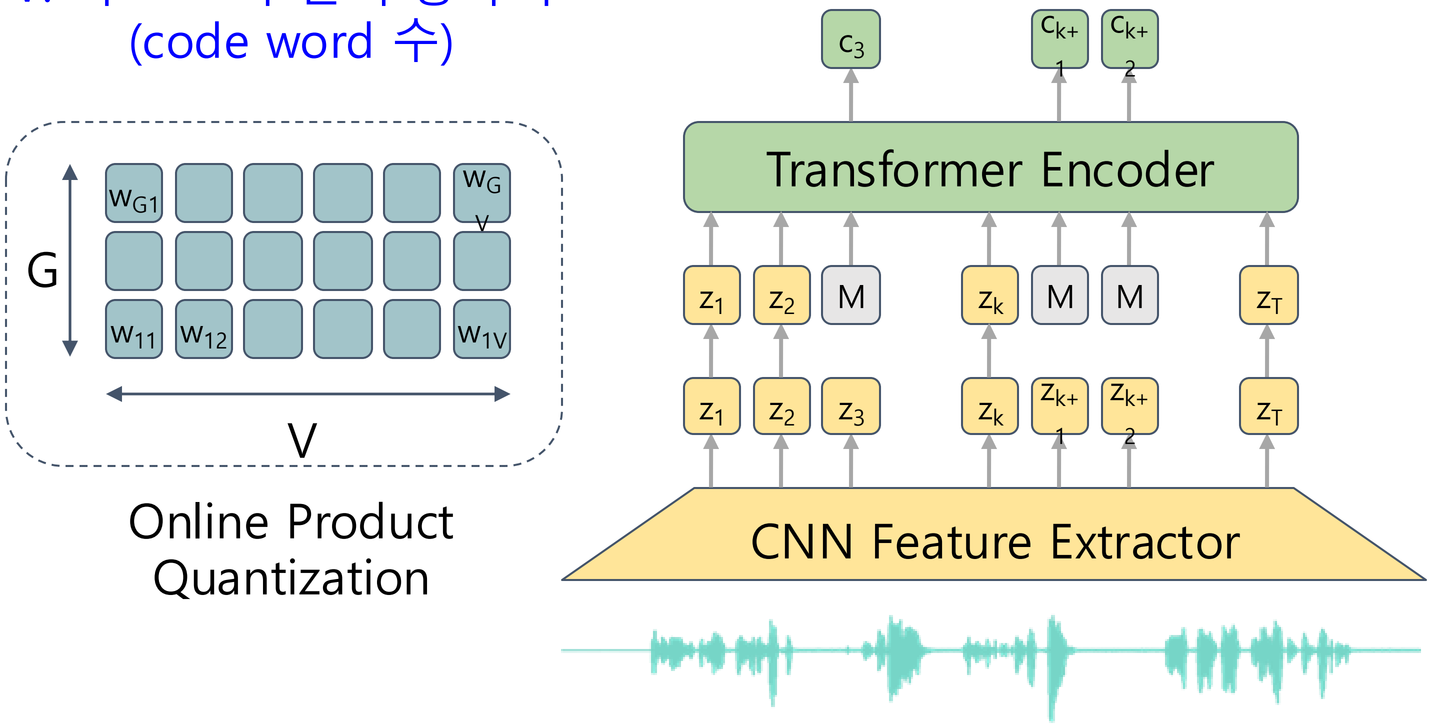


c_3
c_{k+1}
c_{k+2}
Transformer Encoder
z_1
z_2
M
z_k
M
M
z_T
z_1
z_2
z_3
z_k
z_{k+1}
z_{k+2}
z_T
CNN Feature Extractor

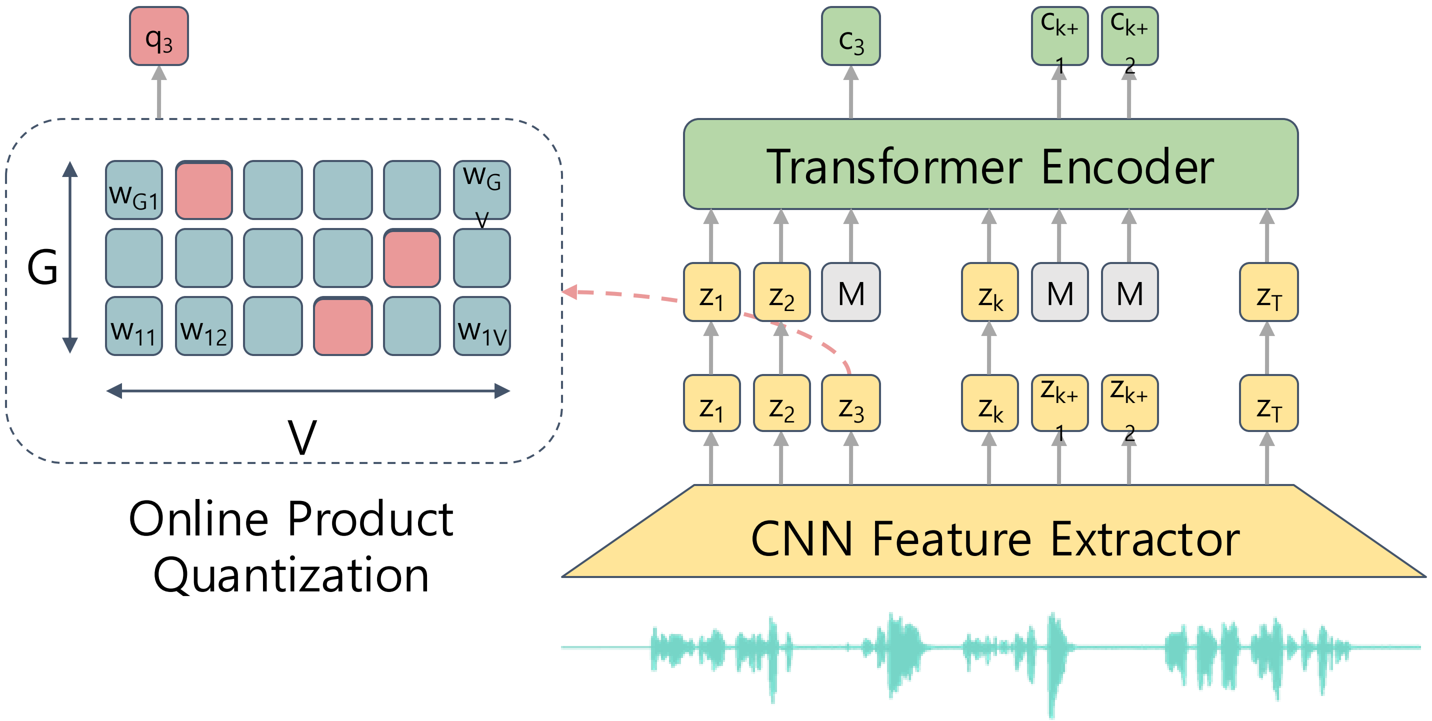
wav2vec2.0

V: 각 코드북 안의 항목 수
(code word 수)

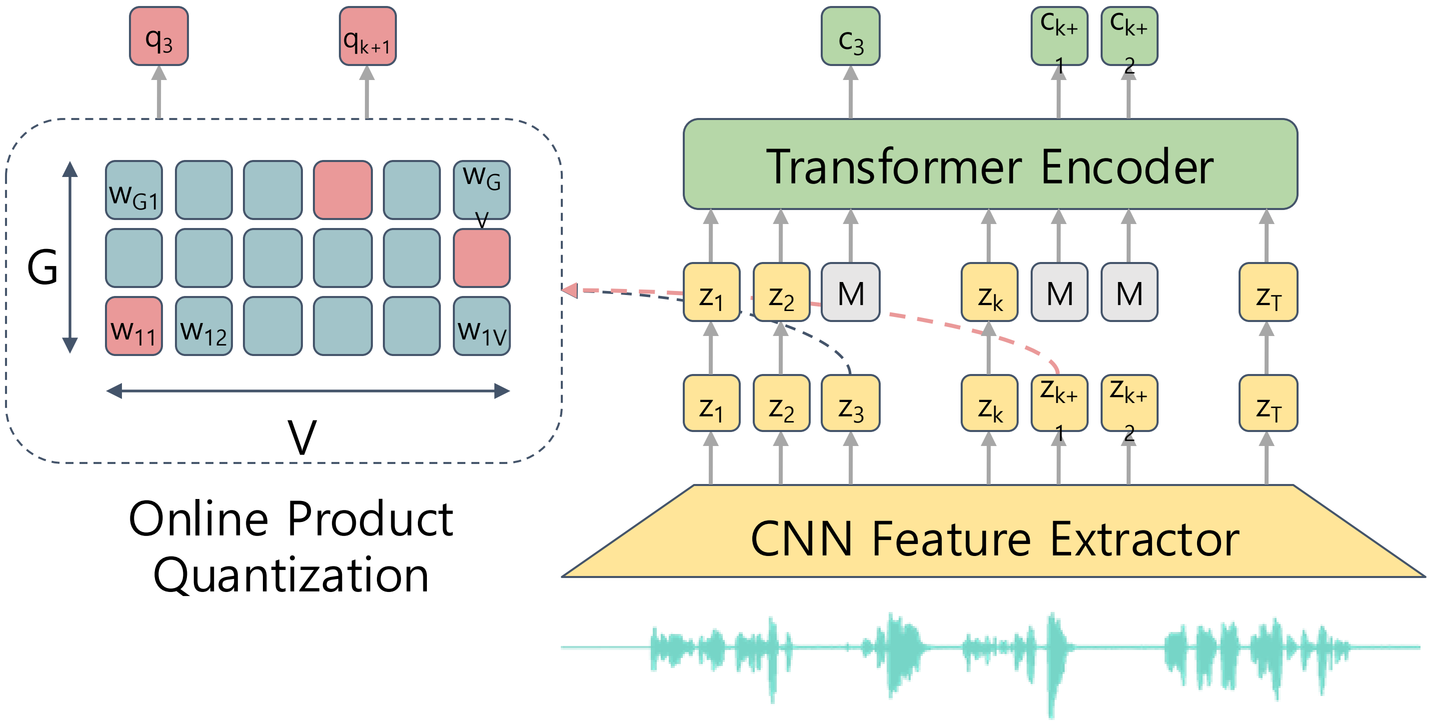
G: 코드북의 개수



wav2vec2.0

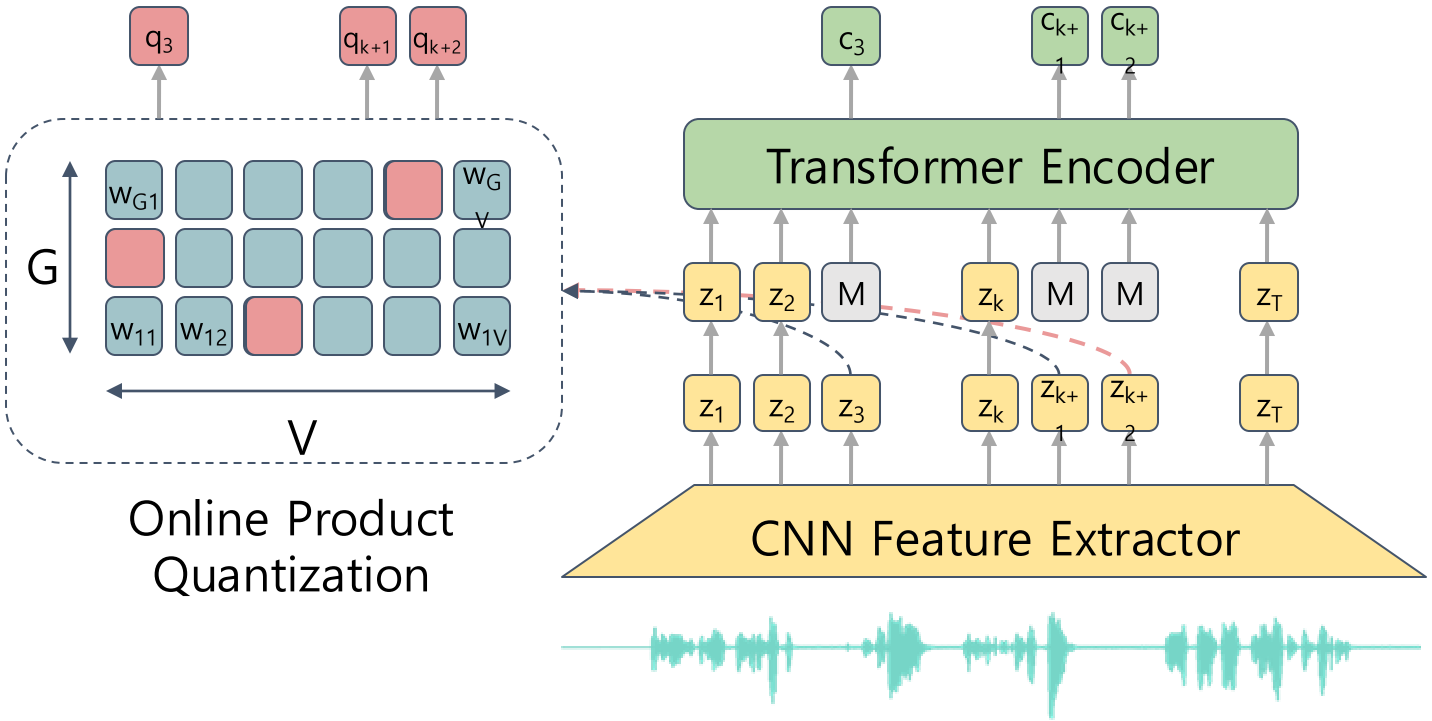

q_3
G
V
Online Product
Quantization
c_3
c_{k+1}
c_{k+2}
Transformer Encoder
z_1 z_2 M z_k M M z_T
z_1 z_2 z_3 z_k z_{k+1} z_{k+2} z_T
CNN Feature Extractor

wav2vec2.0



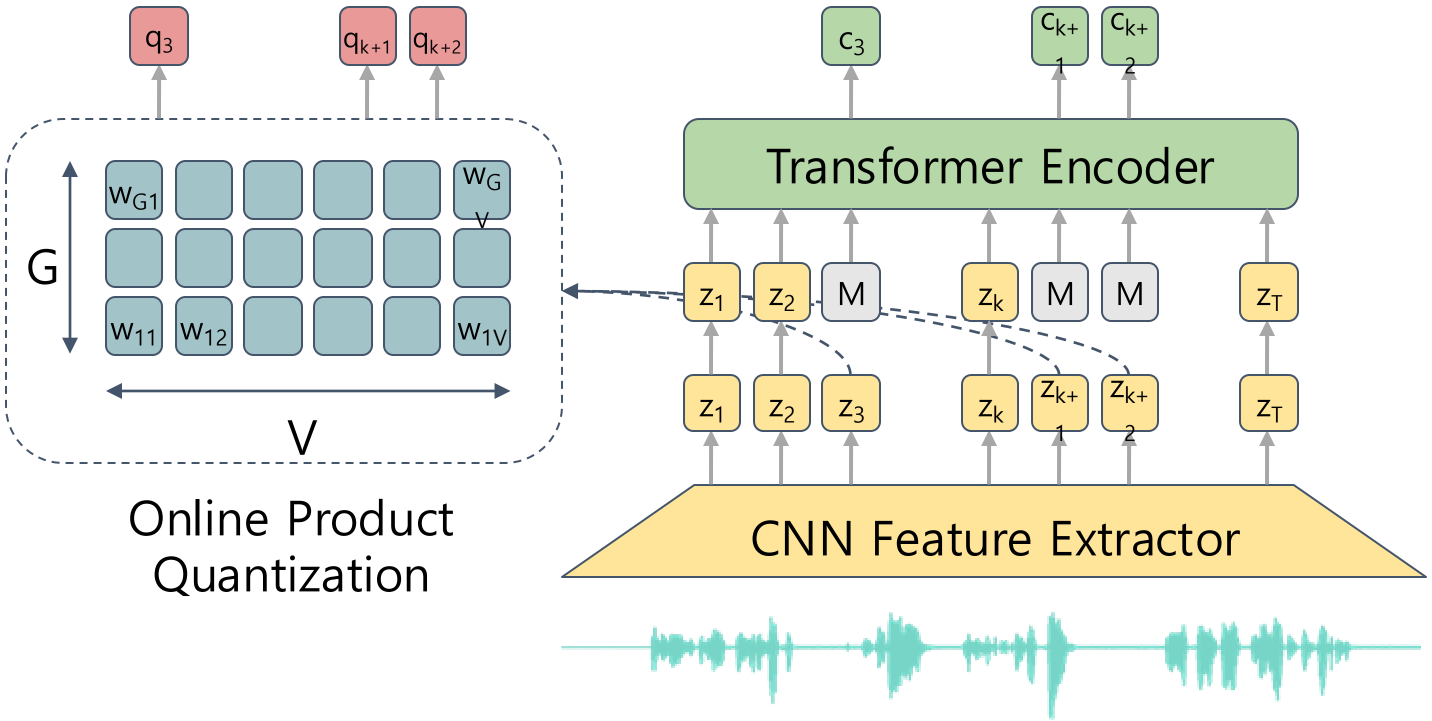
q3
qk+1
c3
ck+1
ck+2
Transformer Encoder
wG1
wGV
G
w11
w12
w1V
V
z1
z2
M
zk
M
M
zT
z1
z2
z3
zk
zk+1
zk+2
zT
CNN Feature Extractor
Online Product Quantization

wav2vec2.0



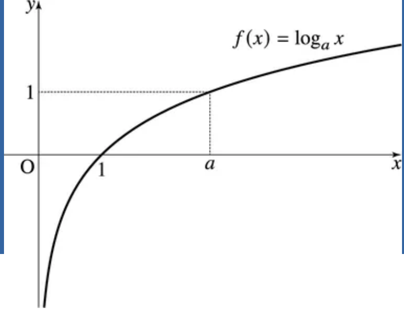
q_3
q_{k+1}
q_{k+2}
c_3
c_{k+1}
c_{k+2}
Transformer Encoder
G
V
w_{G1}
w_{GV}
w_{11}
w_{12}
w_{1V}
z_1
z_2
M
z_k
M
M
z_T
z_1
z_2
z_3
z_k
z_{k+1}
z_{k+2}
z_T
Online Product Quantization
CNN Feature Extractor

wav2vec2.0

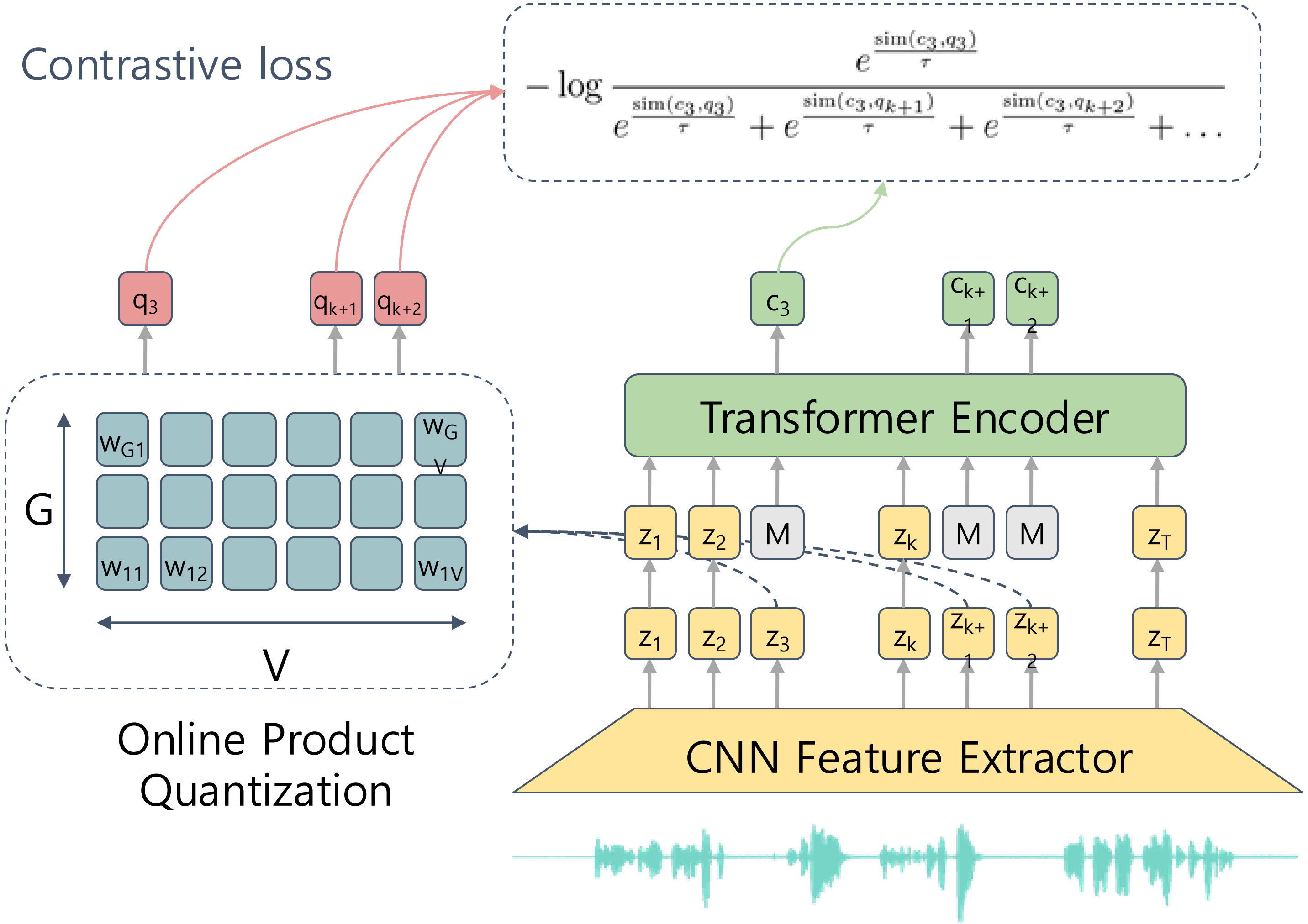


q3
qk+1
qk+2
c3
ck+1
ck+2
Transformer Encoder
wG1
wGV
G
w11
w12
w1V
V
z1
z2
M
zk
M
M
zT
z1
z2
z3
zk
zk+1
zk+2
zT
CNN Feature Extractor
Online Product Quantization

wav2vec2.0

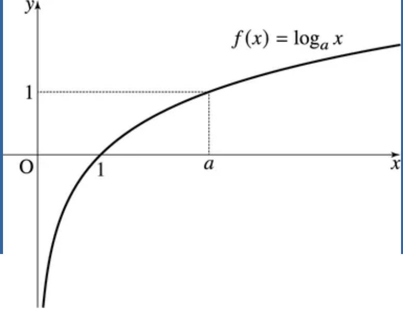


$f(x) = \log_a x$

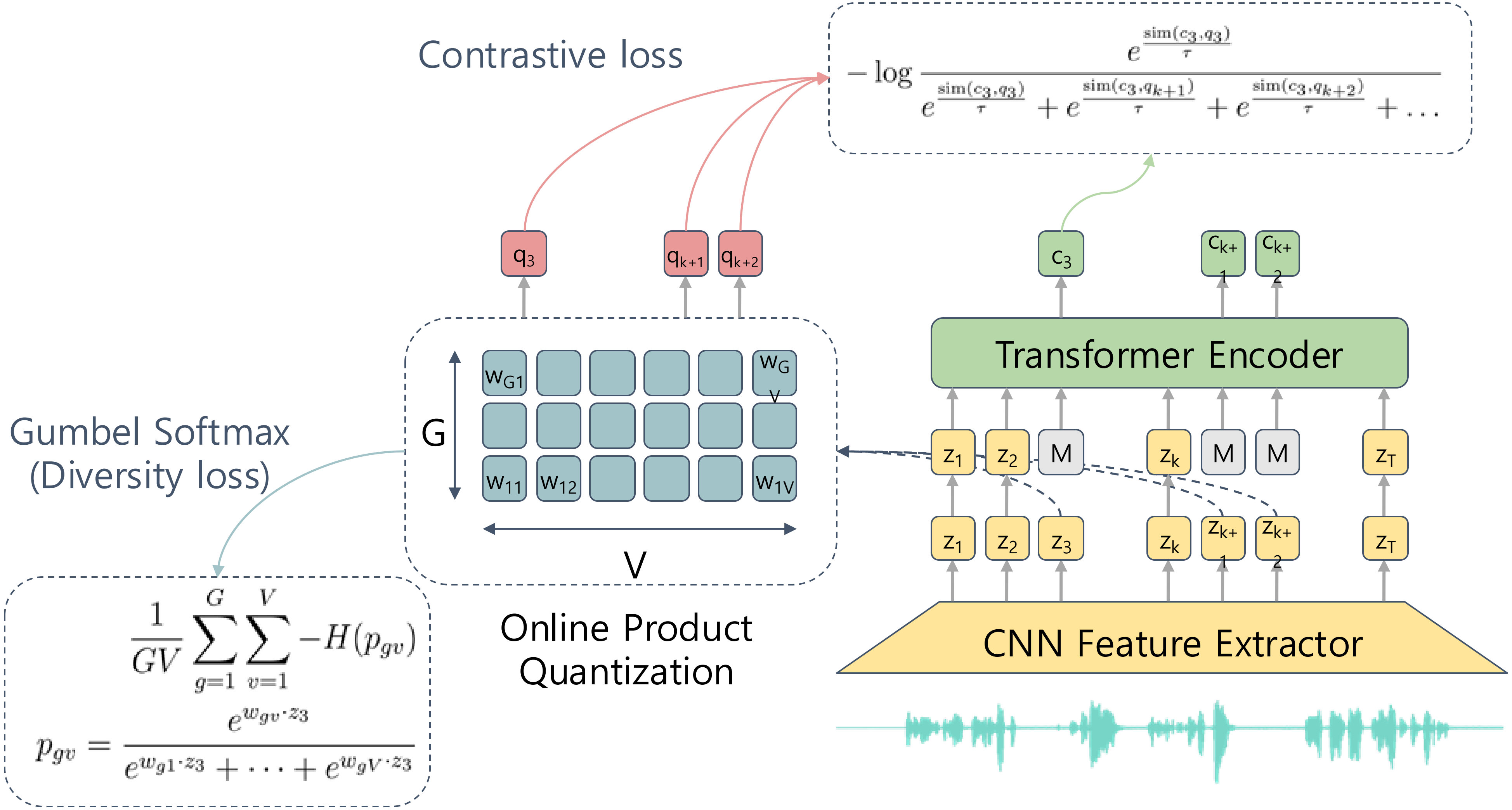


Contrastive loss
$$-\log \frac{e^{\frac{\mathrm{sim}(c_3, q_3)}{\tau}}}{e^{\frac{\mathrm{sim}(c_3, q_3)}{\tau}} + e^{\frac{\mathrm{sim}(c_3, q_{k+1})}{\tau}} + e^{\frac{\mathrm{sim}(c_3, q_{k+2})}{\tau}} + \dots}$$
Transformer Encoder
CNN Feature Extractor
Online Product Quantization

wav2vec2.0



y
f(x) = log_a x
1
O
1
a
x



Contrastive loss
$-\log \frac{e^{\frac{\mathrm{sim}(c_3,q_3)}{\tau}}}{e^{\frac{\mathrm{sim}(c_3,q_3)}{\tau}} + e^{\frac{\mathrm{sim}(c_3,q_{k+1})}{\tau}} + e^{\frac{\mathrm{sim}(c_3,q_{k+2})}{\tau}} + \dots}$
q_3
q_{k+1}
q_{k+2}
c_3
c_{k+1}
c_{k+2}
Transformer Encoder
Gumbel Softmax
(Diversity loss)
w_{G1}
w_{GV}
w_{11}
w_{12}
w_{1V}
G
V
z_1
z_2
M
z_k
M
M
z_T
z_1
z_2
z_3
z_k
z_{k+1}
z_{k+2}
z_T
$\frac{1}{GV}\sum_{g=1}^{G}\sum_{v=1}^{V} -H(p_{gv})$
$p_{gv} = \frac{e^{w_{gv}\cdot z_3}}{e^{w_{g1}\cdot z_3} + \dots + e^{w_{gV}\cdot z_3}}$
Online Product
Quantization
CNN Feature Extractor

wav2vec2.0

- wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations (Facebook AI, 2020)
- K = 100 distractors from the same utterance

Contrastive objective

$$\mathcal{L}_m = -\log \frac{\exp(sim(\mathbf{c}_t, \mathbf{q}_t)/\kappa)}{\sum_{\tilde{\mathbf{q}} \sim \mathbf{Q}_t} \exp(sim(\mathbf{c}_t, \tilde{\mathbf{q}})/\kappa)}$$

$$\mathcal{L} = \mathcal{L}_m + \alpha \mathcal{L}_d$$

Gumbel Softmax

$$\mathcal{L}_d = \frac{1}{GV} \sum_{g=1}^{G} -H(\bar{p}_g) = \frac{1}{GV} \sum_{g=1}^{G} \sum_{v=1}^{V} \bar{p}_{g,v} \log \bar{p}_{g,v}$$

wav2vec2.0

CTC fine-tuning

Base: 96M params

Large: 317M params

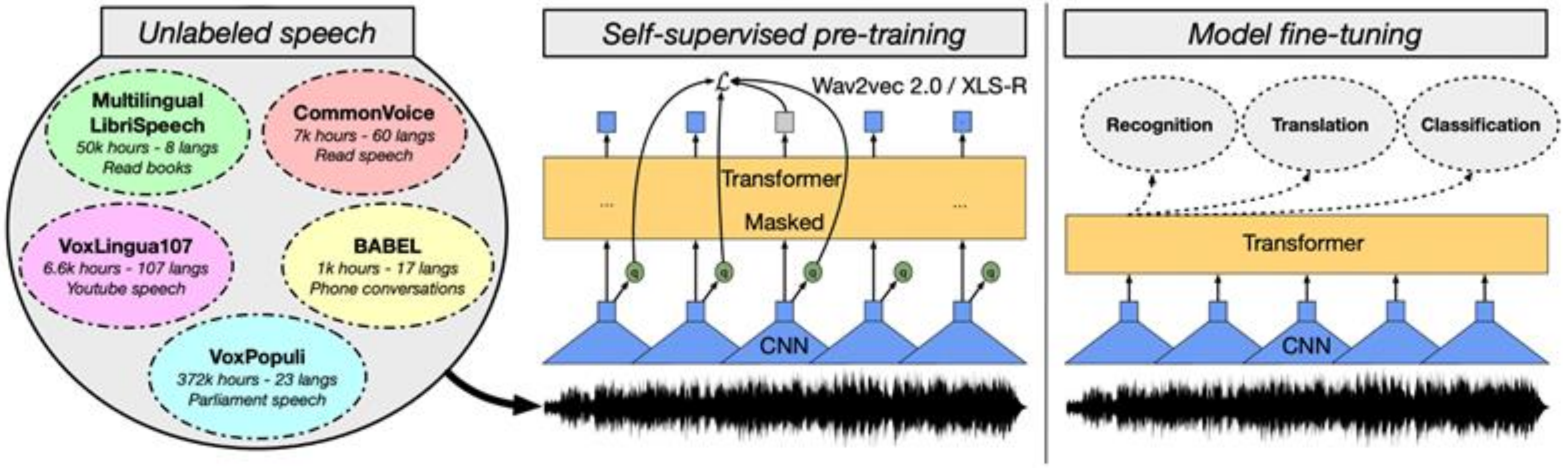
Table 2: WER on Librispeech when using all 960 hours of labeled data (cf. Table 1).

Model	Unlabeled data	LM	dev clean	dev other	test clean	test other
This work						
LARGE - from scratch	-	Transf.	1.7	4.3	2.1	4.6
BASE	LS-960	Transf.	1.8	4.7	2.1	4.8
LARGE	LS-960	Transf.	1.7	3.9	2.0	4.1
	LV-60k	Transf.	1.6	3.0	1.8	3.3

Model	Unlabeled data	LM	dev clean	dev other	test clean	test other
10 min labeled						
Discrete BERT [4]	LS-960	4-gram	15.7	24.1	16.3	25.2
BASE	LS-960	4-gram	8.9	15.7	9.1	15.6
		Transf.	6.6	13.2	6.9	12.9
LARGE	LS-960	Transf.	6.6	10.6	6.8	10.8
	LV-60k	Transf.	4.6	7.9	4.8	8.2
1h labeled						
Discrete BERT [4]	LS-960	4-gram	8.5	16.4	9.0	17.6
BASE	LS-960	4-gram	5.0	10.8	5.5	11.3
		Transf.	3.8	9.0	4.0	9.3
LARGE	LS-960	Transf.	3.8	7.1	3.9	7.6
	LV-60k	Transf.	2.9	5.4	2.9	5.8
10h labeled						
Discrete BERT [4]	LS-960	4-gram	5.3	13.2	5.9	14.1
Iter. pseudo-labeling [58]	LS-960	4-gram+Transf.	23.51	25.48	24.37	26.02
	LV-60k	4-gram+Transf.	17.00	19.34	18.03	19.92
BASE	LS-960	4-gram	3.8	9.1	4.3	9.5
		Transf.	2.9	7.4	3.2	7.8
LARGE	LS-960	Transf.	2.9	5.7	3.2	6.1
	LV-60k	Transf.	2.4	4.8	2.6	4.9
100h labeled						
Hybrid DNN/HMM [34]	-	4-gram	5.0	19.5	5.8	18.6
TTS data augm. [30]	-	LSTM			4.3	13.5
Discrete BERT [4]	LS-960	4-gram	4.0	10.9	4.5	12.1
Iter. pseudo-labeling [58]	LS-860	4-gram+Transf.	4.98	7.97	5.59	8.95
	LV-60k	4-gram+Transf.	3.19	6.14	3.72	7.11
Noisy student [42]	LS-860	LSTM	3.9	8.8	4.2	8.6
BASE	LS-960	4-gram	2.7	7.9	3.4	8.0
		Transf.	2.2	6.3	2.6	6.3
LARGE	LS-960	Transf.	2.1	4.8	2.3	5.0
	LV-60k	Transf.	1.9	4.0	2.0	4.0

wav2vec2.0: XLS-R

- XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale (Meta AI, 2021)
- https://huggingface.co/facebook



Model	dev clean	dev other	test clean	test other
10 min labeled				
wav2vec 2.0 LV-60K (0.3B)	31.7	35.0	32.1	34.5
XLS-R (0.3B)	33.3	39.8	34.1	39.6
XLS-R (1B)	**28.4**	**32.5**	**29.1**	**32.5**
1h labeled				
wav2vec 2.0 LV-60K (0.3B)	13.7	**16.9**	13.7	**17.1**
XLS-R (0.3B)	17.1	23.7	16.8	24.0
XLS-R (1B)	**13.2**	17.0	**13.1**	17.2
10h labeled				
wav2vec 2.0 LV-60K (0.3B)	**5.7**	**9.2**	**5.6**	**9.4**
XLS-R (0.3B)	8.3	15.1	8.3	15.4
XLS-R (1B)	5.9	10.5	5.9	10.6

wav2vec2.0: layer-wise analysis

- [Comparative layer-wise analysis of self-supervised speech models](#) (2022)
- Do we need the raw waveform? ⇒ Spectrograms

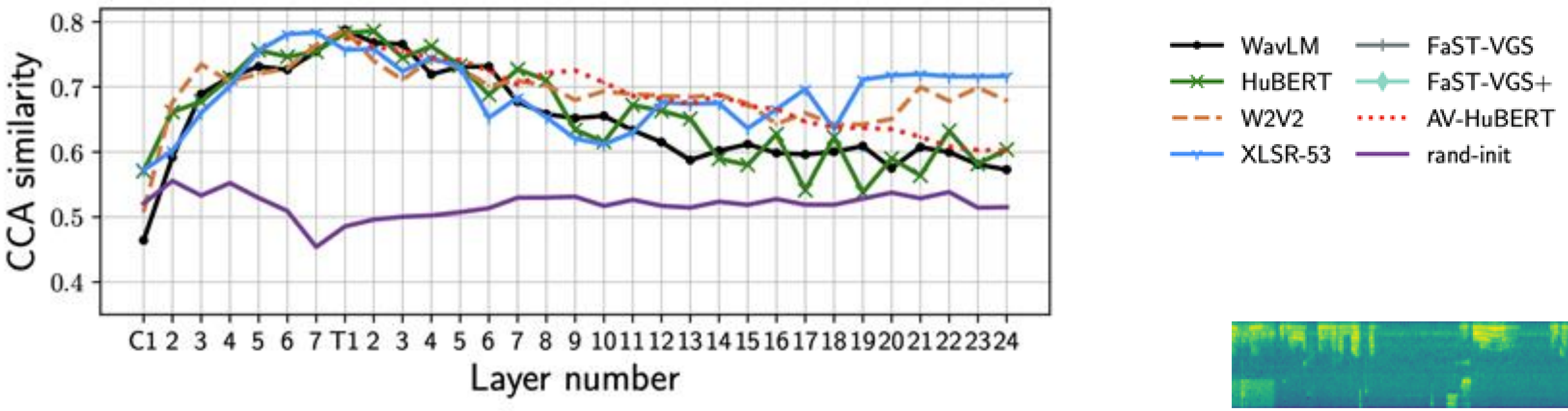
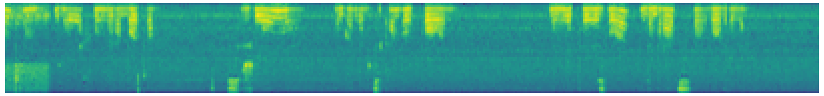
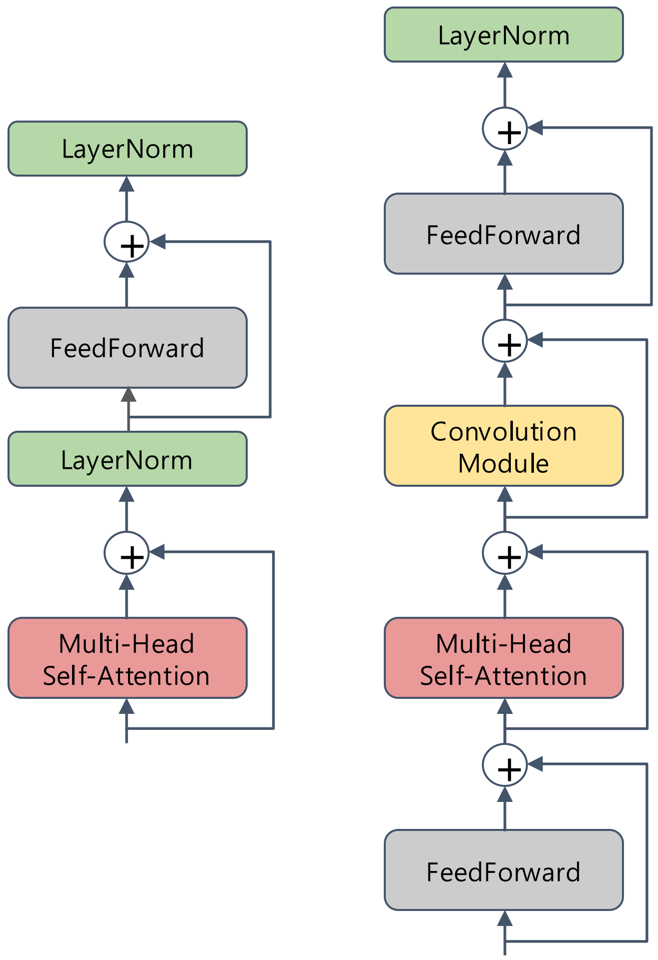


Fig. 2. *CCA similarity with spectrogram features. C_i: CNN layer i, T_j: Transformer layer j.*



Spectrogram example

Conformer



[Conformer: Convolution-augmented Transformer for Speech Recognition]() (Google, 2020)

Table 3: ***Disentangling Conformer.*** *Starting from a Conformer block, we remove its features and move towards a vanilla Transformer block: (1) replacing SWISH with ReLU; (2) removing the convolution sub-block; (3) replacing the Macaron-style FFN pairs with a single FFN; (4) replacing self-attention with relative positional embedding [20] with a vanilla self-attention layer [6]. All ablation study results are evaluated without the external LM.*

Model Architecture	dev clean	dev other	test clean	test other
Conformer Model	1.9	4.4	2.1	4.3
– SWISH + ReLU	1.9	4.4	2.0	4.5
– **Convolution Block**	2.1	4.8	2.1	4.9
– Macaron FFN	2.1	5.1	2.1	5.0
– Relative Pos. Emb.	2.3	5.8	2.4	5.6

[Wav2Vec-Aug: Improved self-supervised training with limited data]() (Meta, 2022)

Modification	Dev-other WER LS-Lab-1H	Dev-other WER LS-Lab-100H
Wav2vec 2.0	12.17	8.17
+ Conformer	**11.34**	**7.32**

wav2vec-Conformer

- Pushing the Limits of Semi-Supervised Learning for Automatic Speech Recognition (Google, 2020)
- LibriVox (60k hours) + LibriSpeech (**960 hours**)

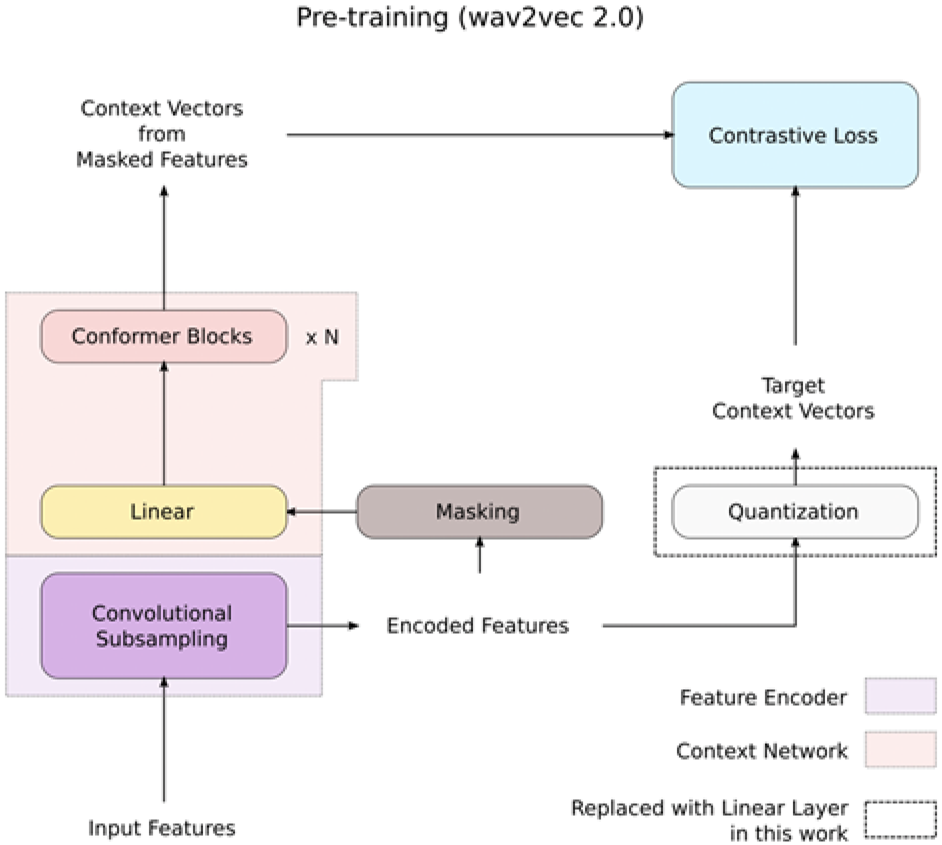


Table 3: WERs(%) from LibriSpeech experiments. LM fusion has not been used.

Method	# Params (B)	dev	dev-other	test	test-other
Trained from scratch					
Conformer L (no rel. attn.)	0.1	2.0	4.7	2.2	4.8
Conformer XL	0.6	2.1	4.9	2.3	4.9
Conformer XLL	1.0	2.3	5.5	2.6	5.6
With pre-training					
Pre-trained Conformer L (no rel. attn.)	0.1	2.0	4.6	2.0	4.5
Pre-trained Conformer XL	0.6	1.7	3.5	1.7	3.5
Pre-trained Conformer XXL	1.0	1.6	3.2	1.6	3.3

wav2vec: frozen quality, layer-wise correlation

- Contrastive Siamese Network for Semi-Supervised Speech Recognition (Google, 2022)
- Comparative layer-wise analysis of self-supervised speech models (2022)

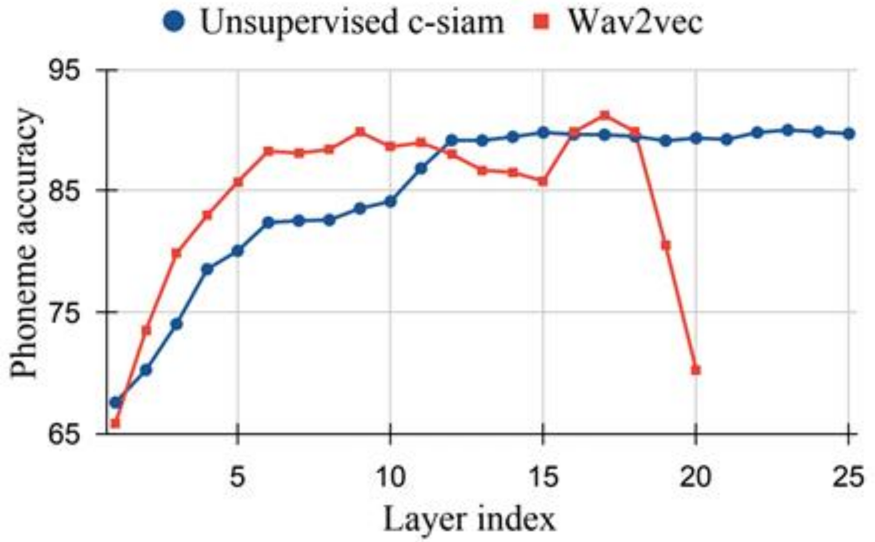


Fig. 2. Frame-level phoneme recognition accuracy using two dense layers for both wav2vec 2.0 (red, square points) and unsupervised c-siam (blue, circular points).

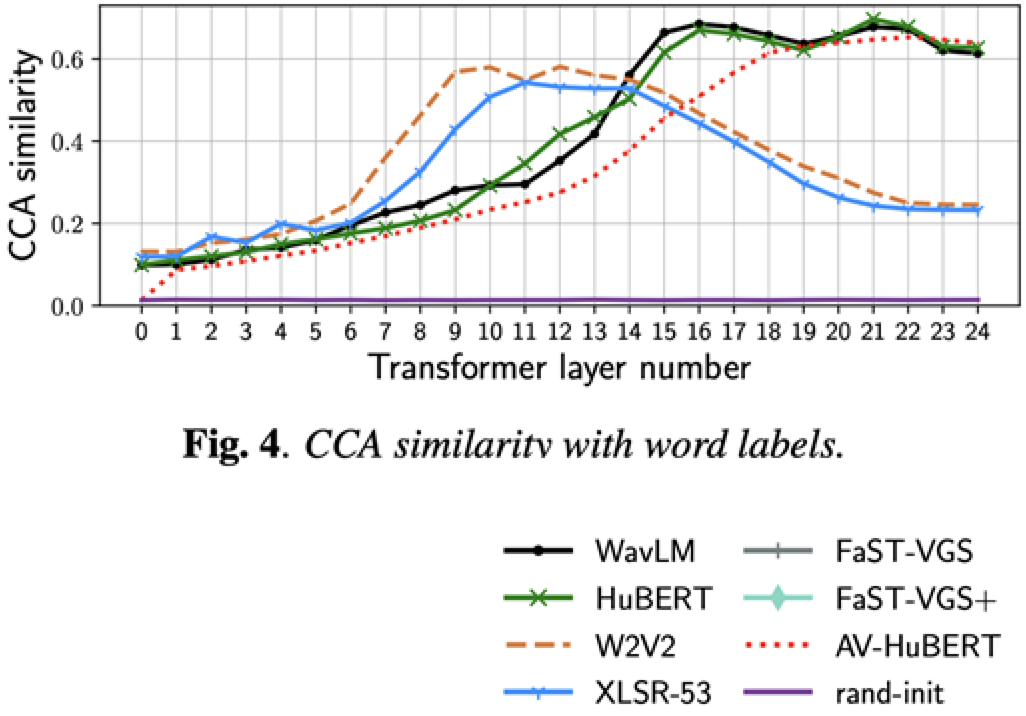
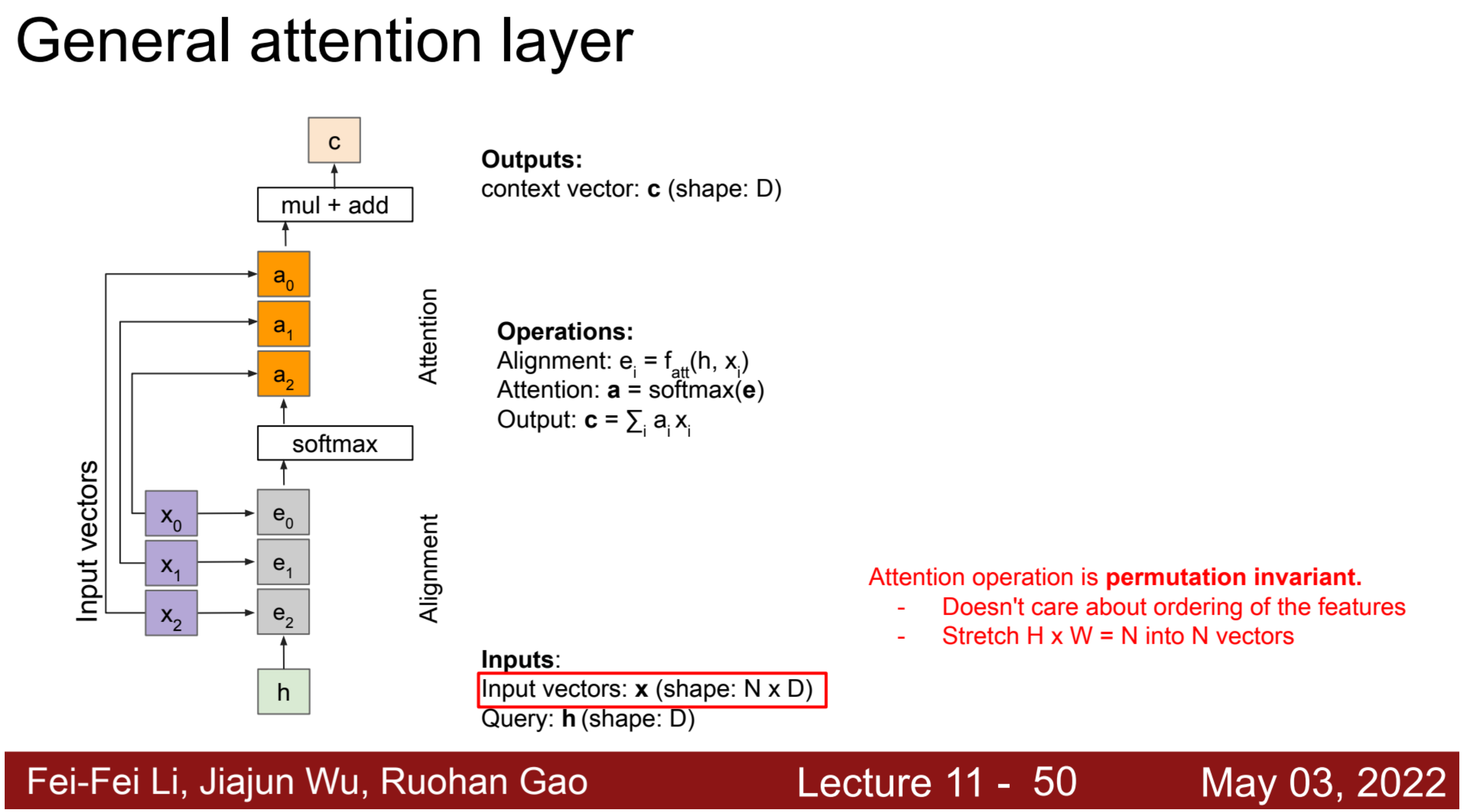


Fig. 4. *CCA similarity with word labels.*

Wav2vec2.0: Summary

- contrastive approach
- The first method to get into single-digit WER on Librispeech test-other using only 10 mins of labels.
- Learnable Vocab: Gumbel Softmax, Diversity Loss
- in-domain pre-training
- waveform → spectrogram
- Transformer → Conformer
- Frozen Encoder ⇒ Autoencoder-like behavior

wav2vec2.0



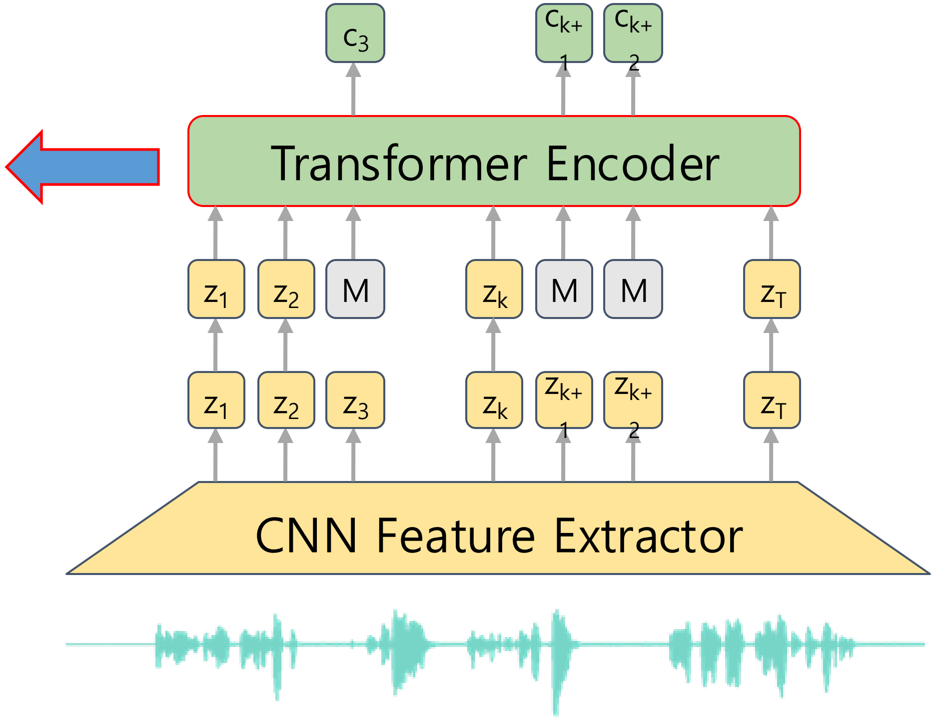
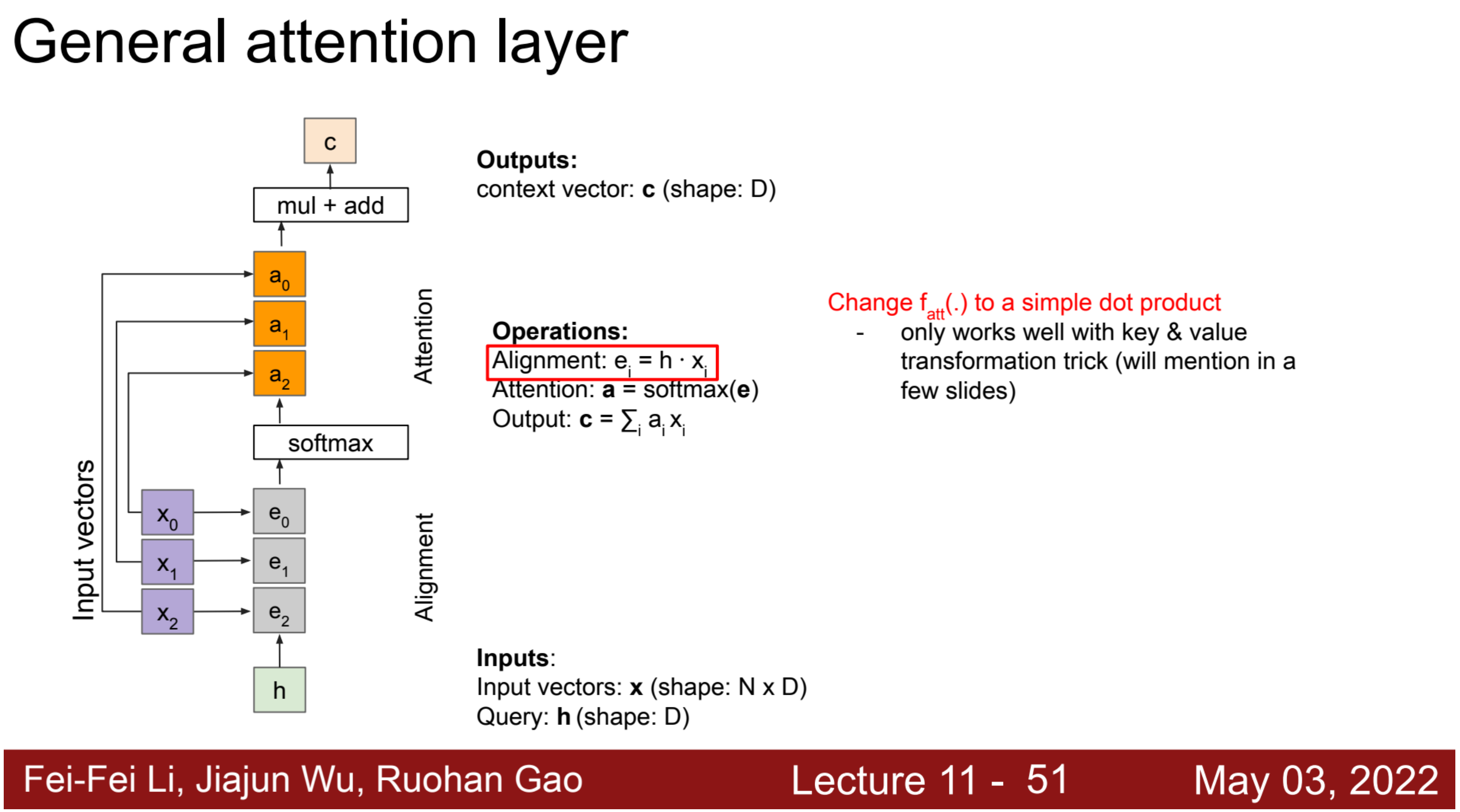


Image from Stanford

wav2vec2.0



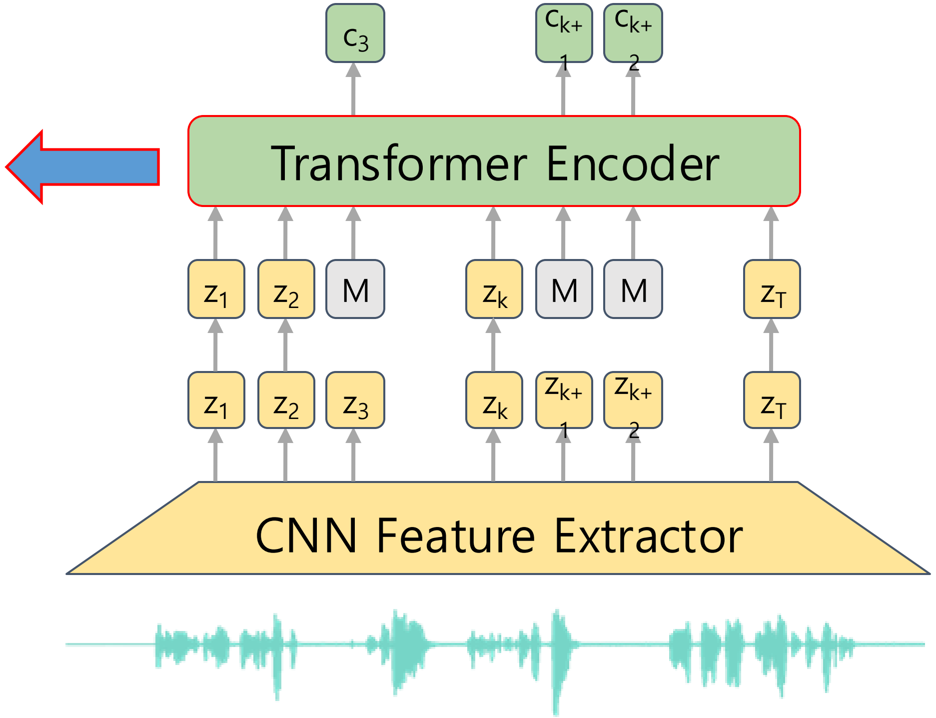
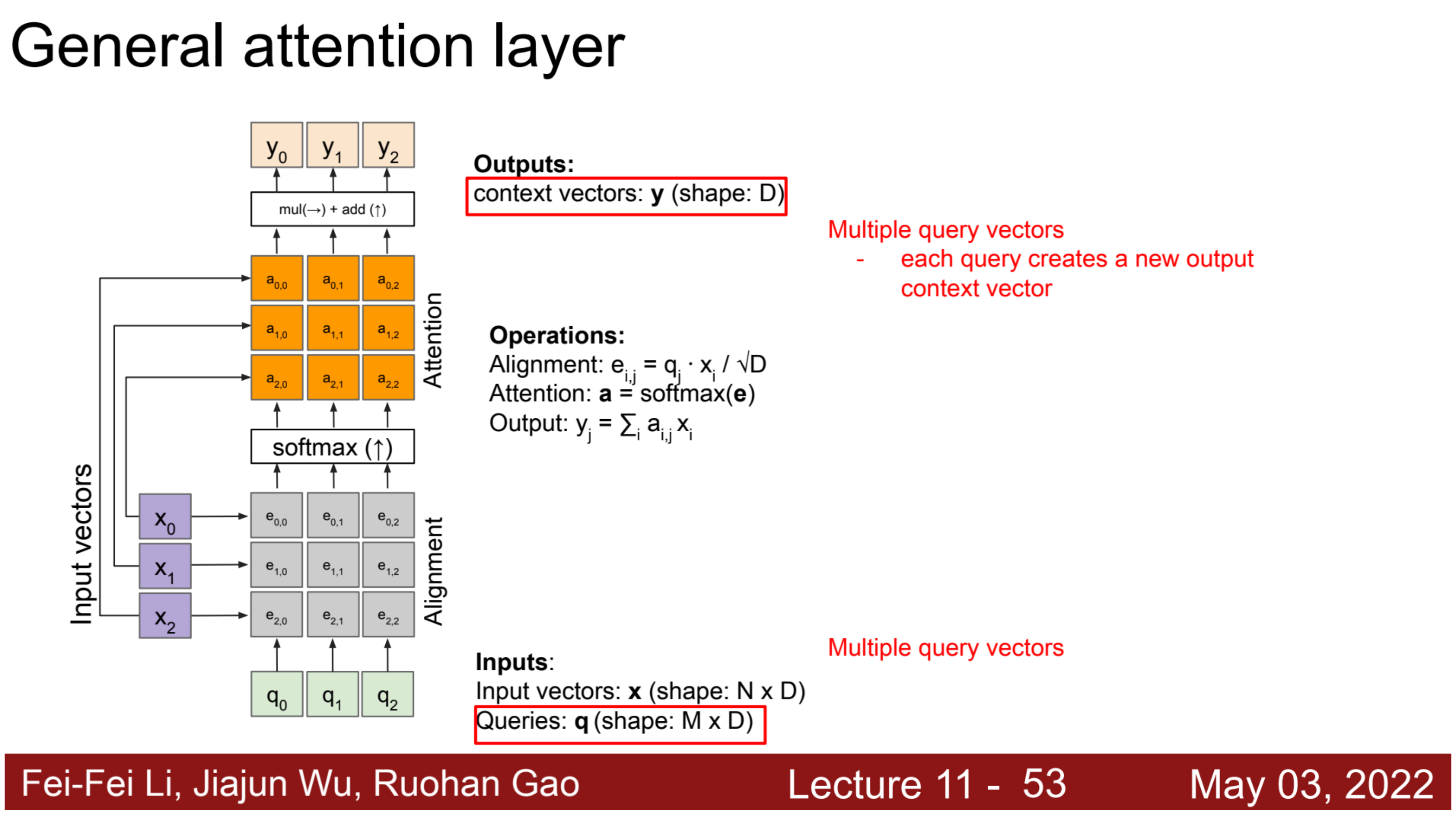


Image from Stanford

wav2vec2.0



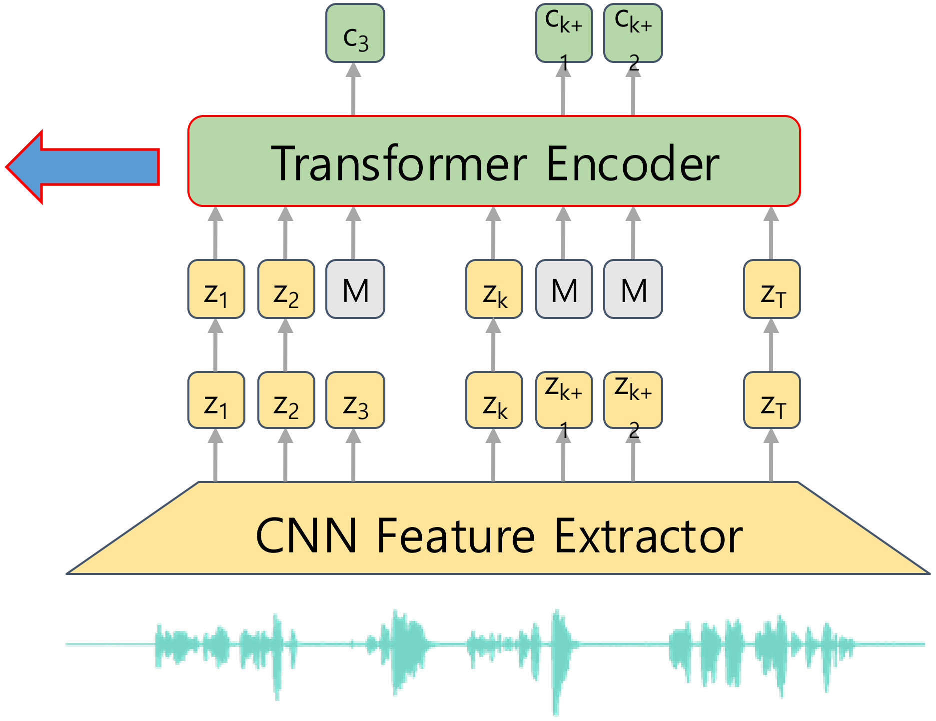
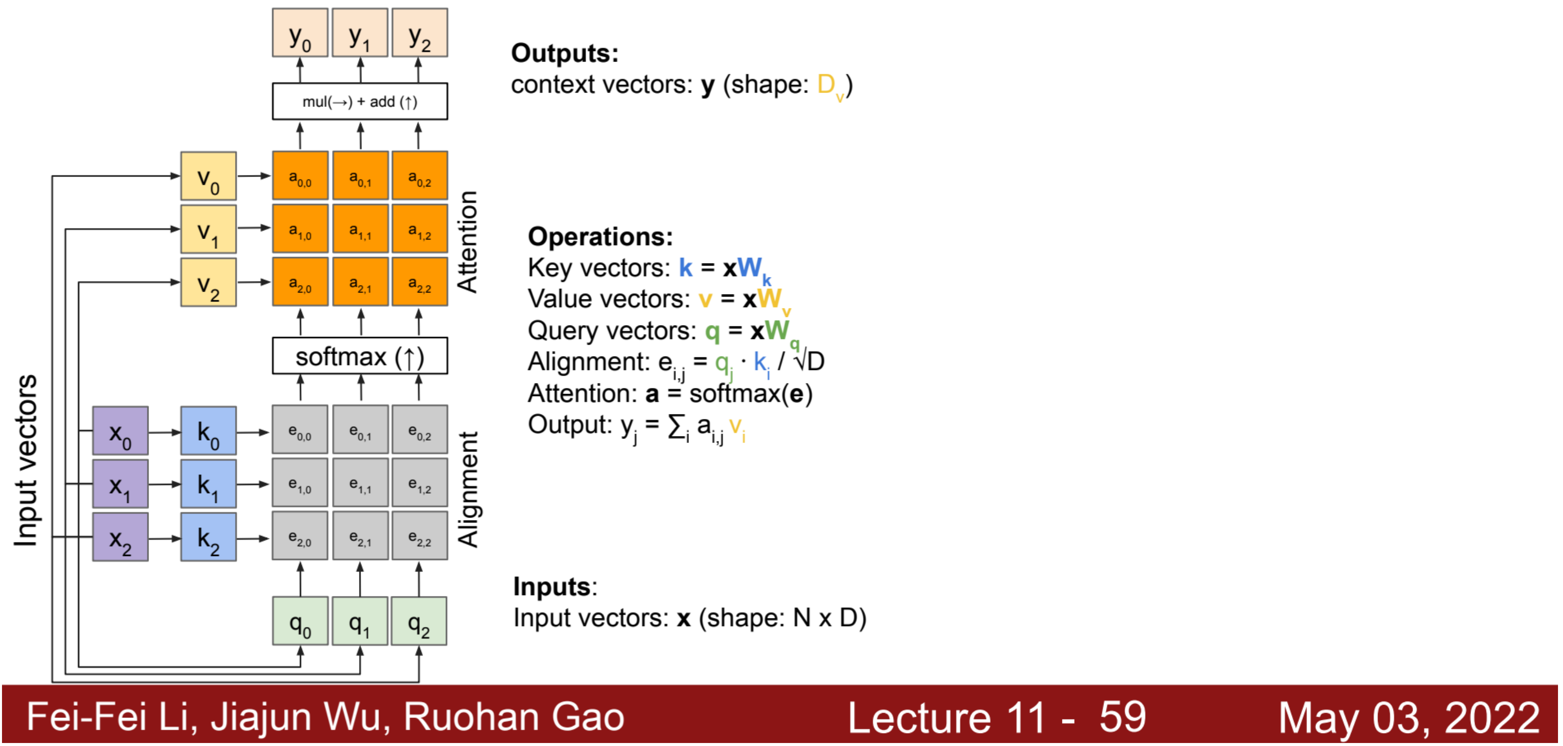
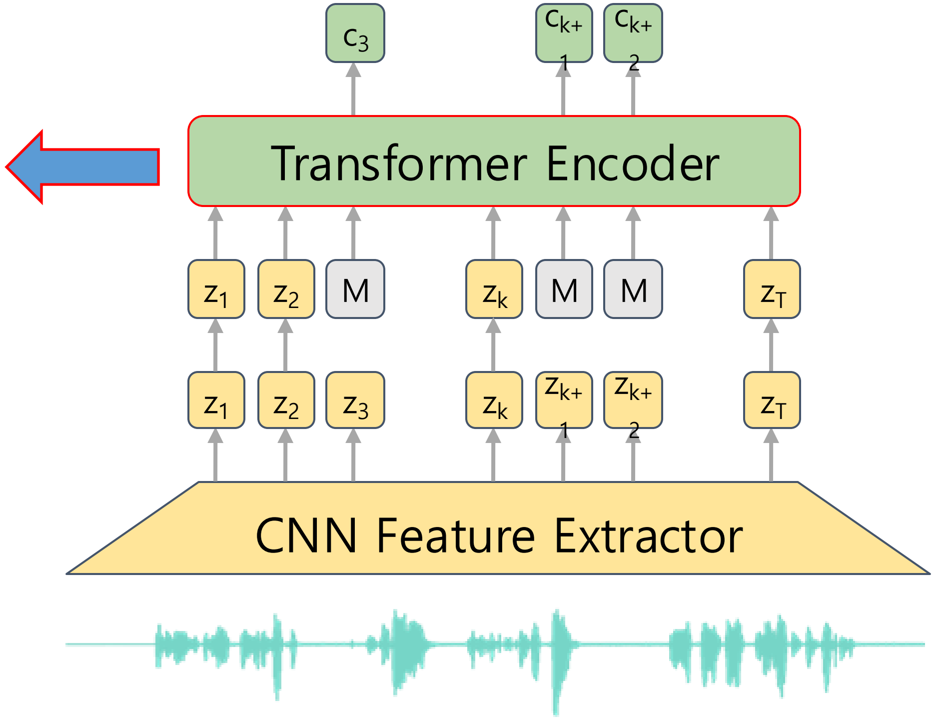


Image from Stanford

wav2vec2.0

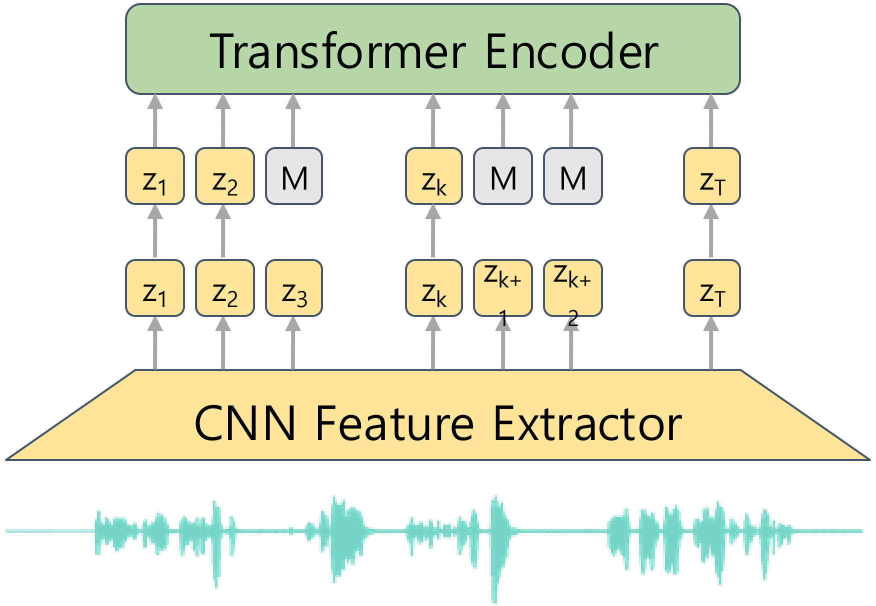
Self attention layer





HuBERT

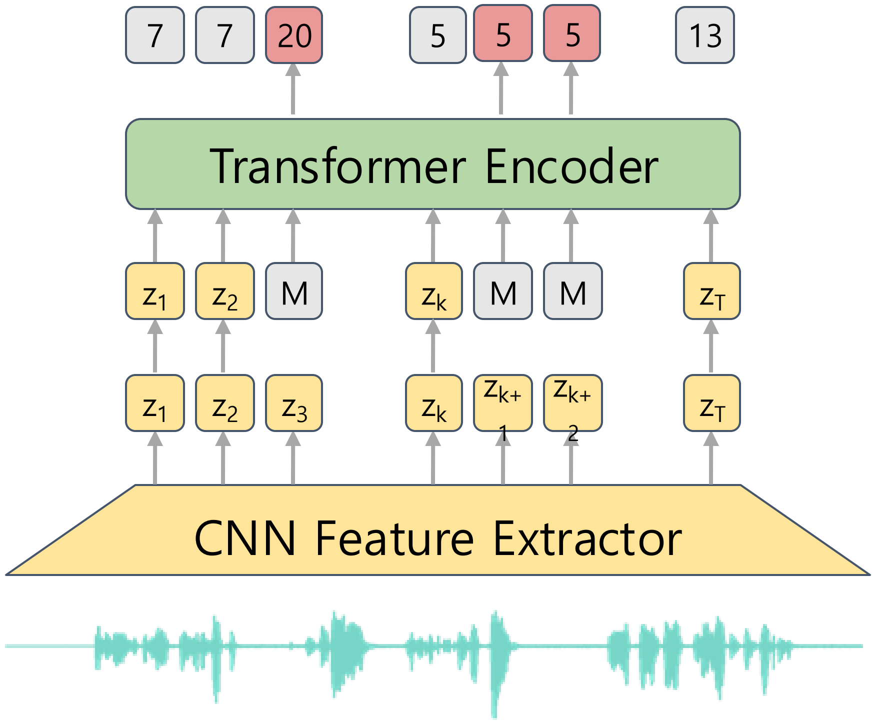
- HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units (Meta, 2021)
- **H**idden **U**nit **BERT**



HuBERT

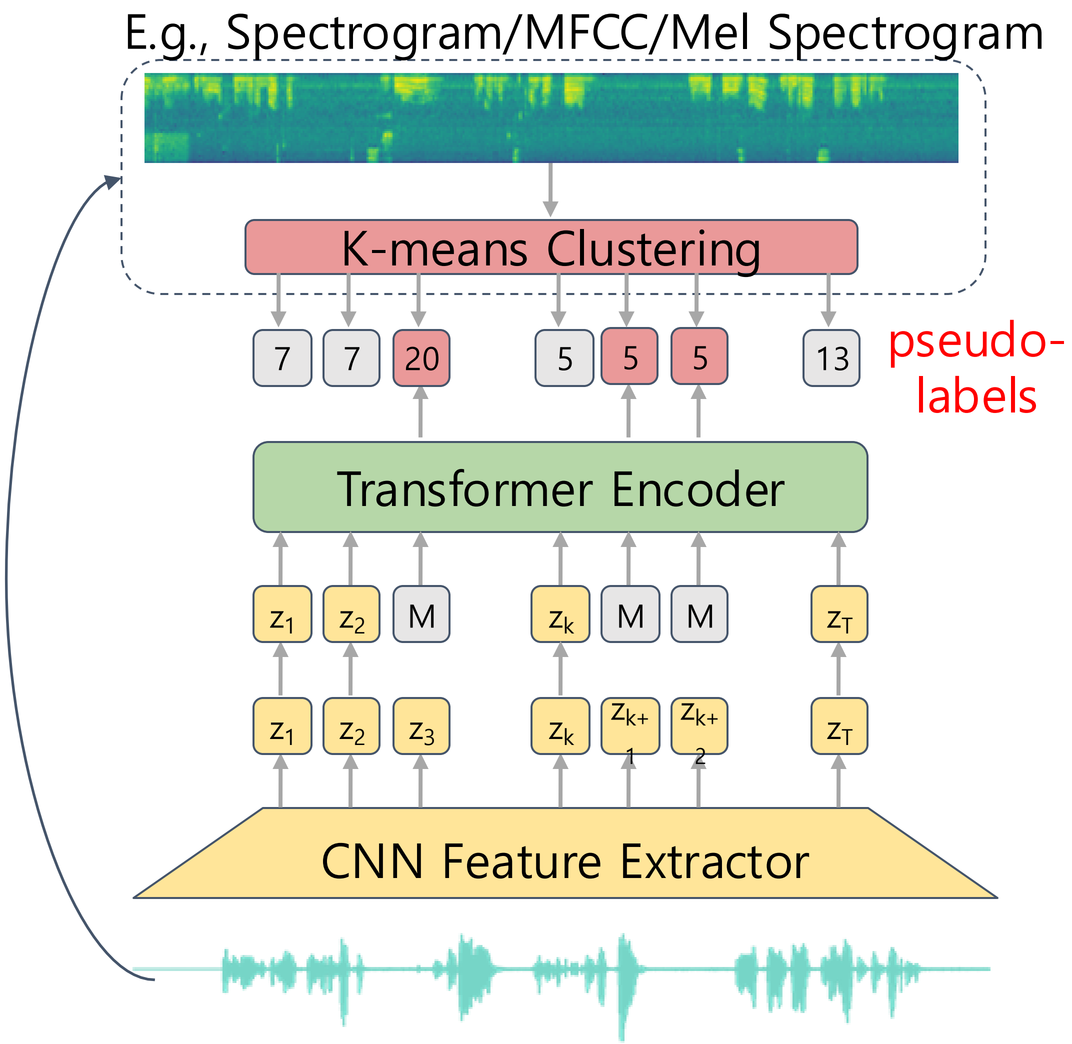
- HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units (Meta, 2021)

- **H**idden **U**nit **BERT**



HuBERT

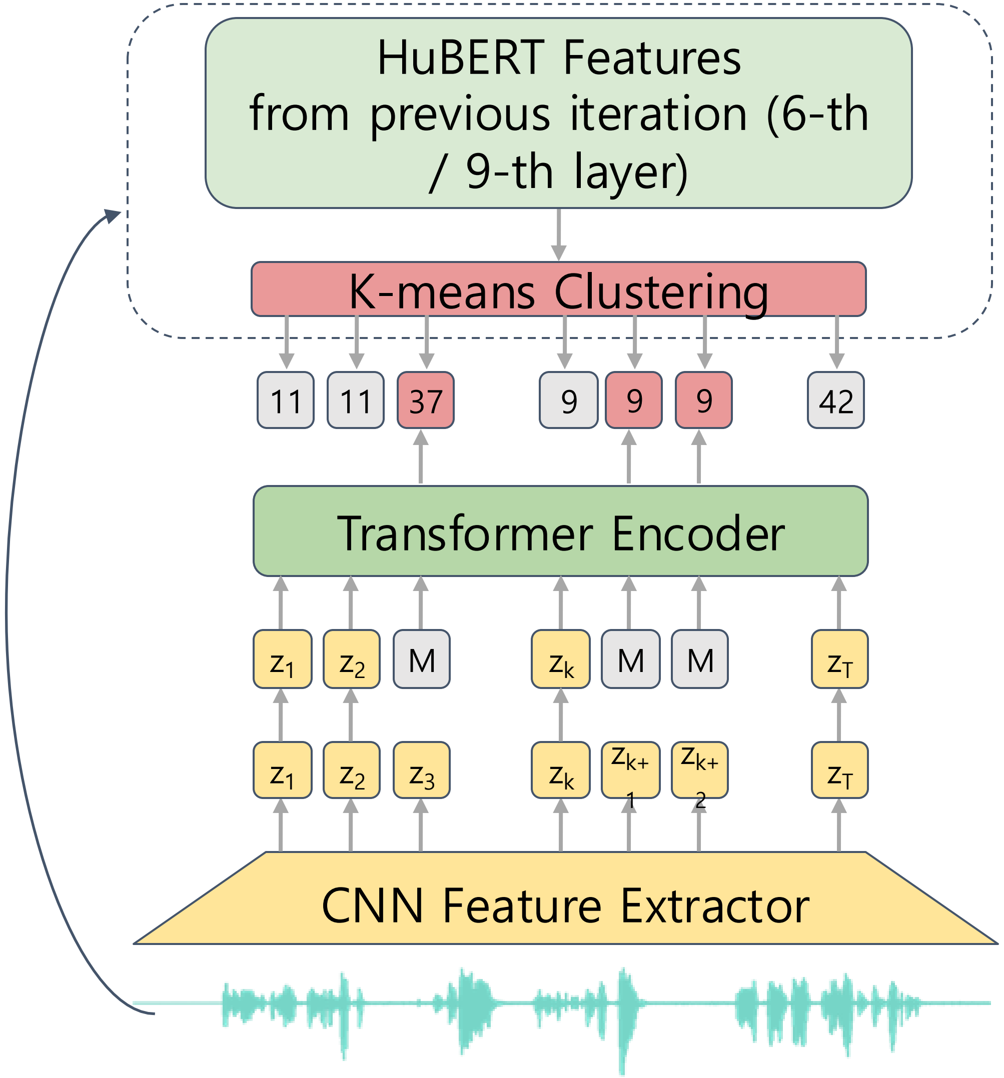
- HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units (Meta, 2021)
- **H**idden **U**nit **BERT**



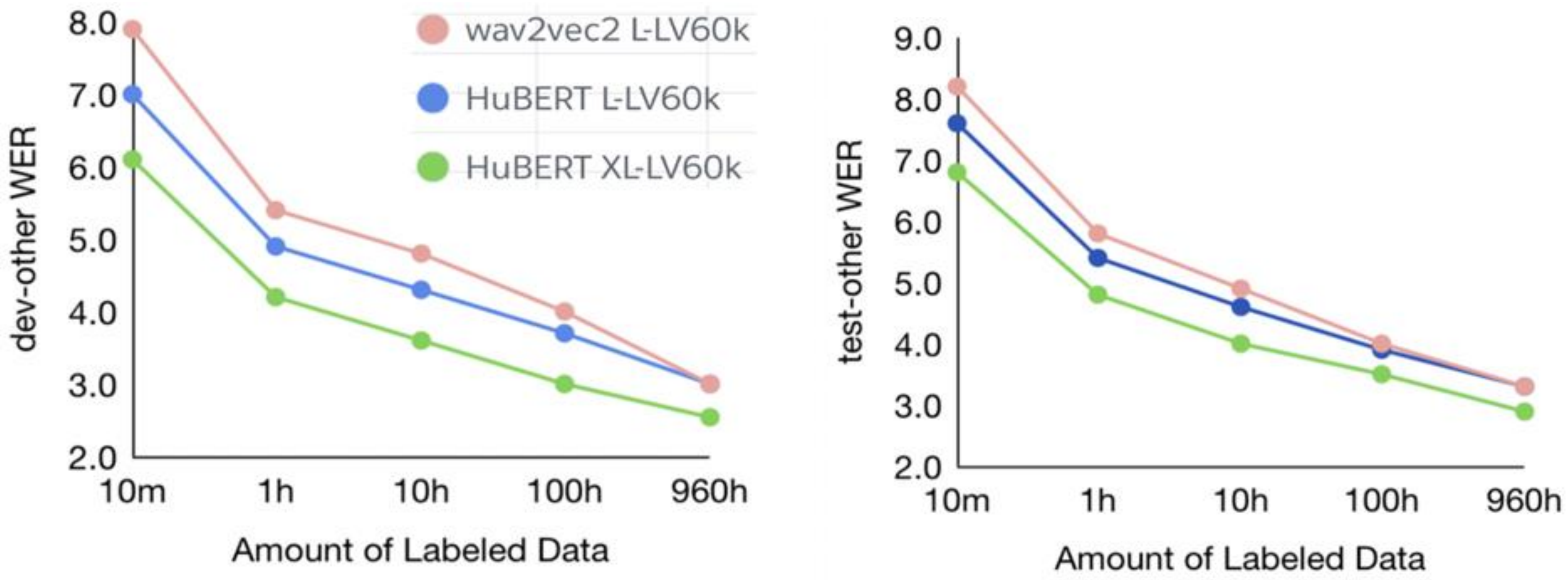
HuBERT

- HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units (Meta, 2021)

- **H**idden **U**nit **BERT**



HuBERT


wav2vec2 L-LV60k
HuBERT L-LV60k
HuBERT XL-LV60k
dev-other WER
8.0
7.0
6.0
5.0
4.0
3.0
2.0
10m
1h
10h
100h
960h
Amount of Labeled Data
test-other WER
9.0
8.0
7.0
6.0
5.0
4.0
3.0
2.0
10m
1h
10h
100h
960h
Amount of Labeled Data

HuBERT

- predictive approach
- no codebook collapse (offline vocab)
- intermediate layer activations as targets